Ατομική Διπλωματική Εργασία

ΑΝΑΚΑΛΥΨΗ ΓΝΩΣΗΣ ΜΕ ΤΗ ΧΡΗΣΗ ΕΞΕΙΔΙΚΕΥΜΕΝΩΝ ΕΡΓΑΛΕΙΩΝ ΛΟΓΙΣΜΙΚΟΥ ΓΙΑ ΤΗΝ ΑΝΑΠΑΡΑΣΤΑΣΗ ΟΝΤΟΛΟΓΙΩΝ

Κυριακή Δημητρίου

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάρτιος 2008

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΑΝΑΚΑΛΥΨΗ ΓΝΩΣΗΣ ΜΕ ΤΗ ΧΡΗΣΗ ΕΞΕΙΔΙΚΕΥΜΕΝΩΝ ΕΡΓΑΛΕΙΩΝ ΛΟΓΙΣΜΙΚΟΥ ΓΙΑ ΤΗΝ ΑΝΑΠΑΡΑΣΤΑΣΗ ΟΝΤΟΛΟΓΙΩΝ

Κυριακή Δημητρίου

Επιβλέπων Καθηγητής
Κωνσταντίνος Παττίχης

Η Ατομική Διπλωματική Εργασία υποβλήθηκε προς μερική εκπλήρωση των απαιτήσεων απόκτησης του πτυχίου Πληροφορικής του Τμήματος Πληροφορικής του Πανεπιστημίου Κύπρου

Μάρτιος 200

Ευχαριστίες

Φτάνοντας στο τέλος της διεκπεραίωσης της ατομικής διπλωματικής μου εργασίας εκτός από τη χαρά και την ικανοποίηση που νιώθω διότι ασχολήθηκα με μια νέα κατεύθυνση στον κλάδο μου που μου πρόσφερε αρκετά, τόσο στο επίπεδο της γνώσης αλλά ήταν και κάτι το ιδιαίτερο, ένα θέμα που αφήνεσαι να το ερευνήσεις και όσο πιο πολύ μαθαίνεις για αυτό, τόσο πιο πολύ συνεχίζεις το ψάξιμο, την έρευνα.

Έτσι ήταν και τα συναισθήματα μου από την αρχή της φοίτησης μου σε αυτό το κλάδο. Μια ακατάπαυστη δίψα για μάθηση σε νέα και συναρπαστικά πράγματα που δεν είχα ξαναδεί ποτέ στη ζωή μου. Μπορεί οι σπουδές να τελειώνουν εδώ, ίσως, όμως η δίψα δεν τελειώνει εδώ, και αυτό διότι η Πληροφορική είναι ένας κλάδος που συνέχεια έχει να σου προσφέρει κάτι καινούργιο να μάθεις.

Αυτό το συναίσθημα όμως πιστεύω το νιώθω χάριν σε όλους αυτούς που με δίδαξαν σε όλη τη διάρκεια των σπουδών μου, τους ακαδημαϊκούς μου. Σημαντικότατη όμως ήταν και η κατά καιρούς βοήθεια που έλαβα από κάποιους από αυτούς, γι αυτό πρωτίστως ευχαριστώ τον καθηγητή με τον οποίο συνεργάστηκα σε αυτή τη μελέτη, τον κ. Παττίχη Κωνσταντίνο.

Τέλος ευχαριστώ την οικογένεια μου και όσους φίλους ήταν κοντά μου κατά τη διάρκεια της προσπάθειας μου αλλά και της φοίτησης μου και με στήριζαν με το δικό τους τρόπο, στους οποίους και αφιερώνω αυτή την εργασία.
Σας ευχαριστώ.

Περίληψη

Η Πληροφορική είναι πλέον ένας από τους πιο εξελίξιμους κλάδους στη σημερινή εποχή της τεχνολογίας και τη βλέπουμε να βρίσκει πάμπολλες εφαρμογές στους πιο πολλούς τομείς σχεδόν της ζωής μας. Μια από τις δυνατότητες της, με την οποία θα ασχοληθούμε εκτενέστερα στη μελέτη αυτή είναι η ανάκτηση και επεξεργασία μεγάλου όγκου δεδομένων σε πολύ μικρό χρονικό διάστημα και με μικρό κόστος.

Πιο συγκεκριμένα, στη μελέτη αυτή θα ασχοληθούμε με την ανάλυση ενός συστήματος που ενσωματώνει μεθόδους και αλγόριθμους για εξόρυξη και επεξεργασία δεδομένων. Θα δημιουργήσουμε βάσεις δεδομένων και θα τις εφαρμόσουμε στο σύστημα αυτό ως εισόδους, προκειμένου να έχουμε ως έξοδο Οντολογικούς χάρτες, μέσα από τους οποίους θα έχουμε ανακάλυψη γνώσης ή κάποιο συσχετισμό και αλληλεξαρτήσεις μεταξύ των δεδομένων της βάσης δεδομένων, τα οποία ήταν προηγουμένως άγνωστα.

Η χρήση του πιο πάνω εργαλείου προς δημιουργία οντολογιών από βιοιατρικά δεδομένα κυρίως όπως θα τη δούμε, θεωρείται πολύ σημαντική και χρήσιμη διότι μπορεί να οδηγήσει στην ανακάλυψη σημαντικής αλλά προηγουμένως άγνωστης πληροφορίας που να σχετίζεται με το θέμα που ερευνάται. Με αυτό τον τρόπο τώρα δίνεται η δυνατότητα στους ερευνητές να εξάγουν νέες πληροφορίες από μεγάλο όγκο πληροφοριών, όπου προηγουμένως αυτό το έκαμναν χειροκίνητα σπαταλώντας συνεπώς περισσότερο χρόνο και ίσως να το έκαμναν λιγότερο αποδοτικά.

Περιεχόμενα

Κεφάλαιο 1

ΕΙΣΑΓΩΓΗ

1.1 Σύντομη Αναφορά Στο Τι Θα Αναπτυχθεί

Ο Παγκόσμιος Ιστός, η πολύ κοινή και συνεχώς αυξανόμενη πηγή πληροφοριών, παρέχει ογκώδεις ετερογενείς συλλογές δεδομένων. Σήμερα βλέπουμε να έχει σημειωθεί μια εκρηκτική αύξηση του όγκου των δεδομένων, γεγονός που βασίζεται στην ωρίμανση των τεχνολογιών των Βάσεων Δεδομένων καθώς και στην ανάπτυξη εργαλείων για την αυτόματη συλλογή δεδομένων.

Με την ύπαρξη τεράστιων βιβλιογραφικών βάσεων δεδομένων, είναι συχνό φαινόμενο κατά συνέπεια να περιέχονται σε αυτές ενδιαφέρουσες πληροφορίες οι οποίες μπορεί να υπονοούνται ή να είναι ακόμα και κρυμμένες. Μια τέτοια βάση δεδομένων είναι η MEDLINE, το κύριο συστατικό του PubMed, η οποία είναι η Εθνική Βιβλιοθήκη Ηνωμένων Πολιτειών Ιατρικής βιβλιογραφικής βάσης, όπου καθημερινά ενημερώνεται με μεγάλο αριθμό βιοιατρικών άρθρων και επιστημονικών αναφορών.

Συνεπώς υπάρχει άμεση ανάγκη για εξαγωγή της χρήσιμης γνώσης που βρίσκεται συσσωρευμένη σε αυτές τις μεγάλες βάσεις δεδομένων καθώς και σε άλλες αποθήκες πληροφορίας.

Στην συγκεκριμένη εργασία επικεντρωνόμαστε στη βιοιατρική πληροφορία και στην αυτόματη ανακάλυψη γνώσεων από έγγραφα βιοϊατρικού περιεχομένου ελεύθερης γραφής (free-texts) κάνοντας χρήση οντολογιών. Η ιατρική πληροφορία απαρτίζεται από κάθε είδους δεδομένα, κείμενο, ή/και οπτικό-ακουστικό υλικό που περιγράφουν πώς να παραμείνει κανείς υγιής, ή πώς να προλάβει μια ασθένεια, ή πώς να φροντίσει/περιθάλψει μια ασθένεια, ή πώς να λάβει τις σωστές εκείνες αποφάσεις που σχετίζονται με την υγεία και την φροντίδα της, αλλά ταυτόχρονα παρέχουν και λεπτομέρειες σχετικά με προϊόντα του χώρου της υγείας και τις διάφορες υπηρεσίες υγείας. Αν τέτοια πληροφορία έχει παρουσία στο διαδίκτυο ορίζεται ως διαδικτυακή ιατρική πληροφορία.

Σήμερα διάφοροι κυβερνητικοί και μη οργανισμοί και ιδρύματα υγείας, ιατρικές σχολές, εταιρείες δραστηριοποιούμενες στο χώρο της υγείας, ιατρικός περιοδικός τύπος δημοσιεύουν πληροφορία σχετική με βιβλιογραφία, ασθένειες, φάρμακα, εμπορικά ιατρικά προϊόντα και μπορούν να χρησιμοποιηθούν ως κανάλια επικοινωνίας ανάμεσα σε ασθενείς, γιατρούς και ιδρύματα/οργανισμούς. Η αναζήτηση αυτής της πληροφορίας στο διαδίκτυο σήμερα γίνεται με τη χρήση μηχανών αναζήτησης ή τη χρήση πυλών υγείας, όπως το MEDLINE που αναφέρθηκε πιο πάνω.

Ένα όλο και μεγαλύτερο πλήθος από «καταναλωτές» υπηρεσιών υγείας χρησιμοποιεί σήμερα το διαδίκτυο για την διερεύνηση θεμάτων υγείας. Παρόλα αυτά συναντώνται κάποια προβλήματα όσον αφορά την ιατρική πληροφορία, όπως το ότι υπάρχουν πολλά sites με πάρα πολύ πληροφορία(υπερφόρτωση πληροφοριών), μη ενοποιημένο και ομογενοποιημένο περιβάλλον και μη φιλτραρισμένη πληροφορία στις περισσότερες περιπτώσεις με κίνδυνο έτσι για «άχρηστη» πληροφορία. Συνεπώς οδηγούμαστε στην επιτακτική ανάγκη για εξόρυξη δεδομένων για εξαγωγή της χρήσιμης μόνο πληροφορίας ή καλύτερα της πληροφορίας που μας ενδιαφέρει.

Τεχνικές εξόρυξης δεδομένων από κείμενα, κατασκευή οντοτήτων και λεπτομερής ανάλυση συγκεκριμένου λογισμικού που καταπιάνεται με αυτές τις έννοιες θα δούμε παρακάτω.

1.2 Στόχοι Διπλωματικής Εργασίας

Η μελέτη αυτή ασχολείται με τη μελέτη και ανάλυση ενός συστήματος δημιουργίας Οντολογιών. Πρόκειται για το σύστημα "OntoGen", το οποίο συνδυάζει τεχνικές εξόρυξης κειμένων με μια αποδοτική αλληλεπίδραση με το χρήστη, έτσι ώστε να μειώσει το χρόνο που σπαταλάτε υπό άλλες συνθήκες και την πολυπλοκότητα για το χρήστη.

Στόχος της μελέτης μέσα από το εργαλείο και συγκεκριμένα μέσα από εφαρμογές που θα υλοποιήσουμε, είναι να δείξουμε πως μπορούμε να ανακαλύψουμε γνώση, πληροφορίες από δεδομένα, η οποία δεν είναι ξεκάθαρη στην αρχική της μορφή αλλά υπονοείται. Κάτω από άλλες συνθήκες, χωρίς δηλαδή να είχαμε στη διάθεση μας ένα τέτοιο εργαλείο, η ανακάλυψη αυτή θα ήταν χρονοβόρα και ίσως και αδύνατη.

Ένα σημαντικό εργαλείο όπως θα αποδειχθεί με το τέλος της εργασίας, το οποίο μπορεί να εφαρμοσθεί σε διάφορους τομείς. Από το τομέα της ιατρικής, για την καλύτερη κατανόηση κάποιων ακόμα ανεξήγητων ίσως ασθενειών μέχρι το τομέα του εμπορίου και βιομηχανίας. Ο κάθε ερευνητής μπορεί να το ρυθμίσει στις δικές του προτιμήσεις.

1.3 Ανασκόπηση Διπλωματικής Εργασίας

Στο παρόν κεφάλαιο, κεφάλαιο 1, δίνεται μια σύντομη αναφορά για το τι πρόκειται να αναπτυχθεί στη συνέχεια καθώς και οι στόχοι της μελέτης. Επίσης θα γίνει μια αναφορά για τα επόμενα κεφαλαία που θα ακολουθήσουν.

Στο κεφάλαιο 2, δίνονται κάποιοι βασικοί ορισμοί σχετικοί με την εξόρυξη δεδομένων γενικά και την εξόρυξη δεδομένων στη βιοιατρική περιοχή ειδικότερα, προτού προχωρήσουμε σε θέματα εφαρμογών έτσι ώστε ο αναγνώστης να σχηματίσει μια πρώτη ιδέα για το θέμα που θα δούμε εκτενέστερα στη συνέχεια.

Στο κεφάλαιο 3, γίνεται αναφορά στην έννοια της Οντολογίας και της δημιουργίας αυτής.

Στο κεφάλαιο 4, παρουσιάζεται αναλυτικά το σύστημα “OntoGen” που μελετάμε σε αυτή τη εργασία. Πιο συγκεκριμένα παρουσιάζεται ξεχωριστά κάθε λειτουργία του συστήματος, εξηγώντας κάθε φορά στο χρήστη ποιους κανόνες μάθησης ή τεχνικές εξόρυξης δεδομένων ή αλγόριθμους κατηγοριοποίησης κρύβει από πίσω.

Στο κεφάλαιο 5, αναφέρονται όλοι αυτοί οι αλγόριθμοι και οι μέθοδοι και τεχνικές που υποστηρίζουν τη λειτουργία του συστήματος. Από τα κεφάλαια 4, 5 συμπεραίνουμε ότι το σύστημα ενσωματώνει μεθόδους και αλγόριθμους που ήδη μας είναι γνωστά. Απλώς εμείς μέχρι στιγμής τα χρησιμοποιούσαμε ξεχωριστά το καθένα. Τώρα έχουμε ένα εργαλείο που ενσωματώνει όλες αυτές τις τεχνικές μαζί, πετυχαίνοντας έτσι κάτι περισσότερο.

Στο 6 παρουσιάζεται βήμα προς βήμα μια υλοποίηση στο σύστημα OntoGen και στη συνέχεια γίνεται μια αξιολόγηση του συστήματος.

Από τα κεφάλαια 4, 5 συμπεραίνουμε ότι τι σύστημα ενσωματώνει μεθόδους και αλγόριθμους που ήδη μας είναι γνωστά. Απλώς εμείς μέχρι στιγμής τα χρησιμοποιούσαμε ξεχωριστά το καθένα. Τώρα έχουμε ένα εργαλείο με όλα αυτά μαζί, πετυχαίνοντας έτσι κάτι περισσότερο. Αυτό το κάτι περισσότερο, είναι αυτό που θα δούμε στα κεφάλαιο 7. Πιο συγκεκριμένα παρουσιάζονται τέσσερις εφαρμογές σε πεδία δεδομένων τα οποία επιλέξαμε εμείς και στη συνέχεια παρουσιάζουμε τα σχετικά αποτελέσματα.

Στο τέλος, στο κεφάλαιο 8, κλείνουμε με κάποια συμπεράσματα από την όλη μελέτη καθώς και κάποιες εισηγήσεις για μελλοντικές επεκτάσεις του εργαλείου.

Κεφάλαιο 2

ΥΠΟΒΑΘΡΟ ΣΤΗΝ ΕΞΟΡΥΞΗ ΚΕΙΜΕΝΩΝ & ΒΙΟΙΑΤΡΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ ΕΞΟΡΥΞΗΣ ΛΟΓΟΤΕΧΝΙΑΣ

2.1 Εισαγωγή

Πριν προχωρήσουμε στην εξέταση και ανάλυση του συστήματος κατασκευής οντοτήτων OntoGen, καλό θα ήταν να δούμε κάποιες από τις βασικές έννοιες που θα μας απασχολήσουν κατά τη διάρκεια της μελέτης και στις οποίες στηρίζεται η όλη διαδικασία της ανακάλυψης γνώσεων από βιοιατρική βιβλιογραφία.

2.2 Εξόρυξη Κειμένου (Text Mining) : ένα περίγραμμα πριν δώσουμε τον όρο και την ανάγκη που οδήγησε σε αυτή την τεχνική.

Η δυναμική ανάπτυξης και οι νέες ανακαλύψεις στις περιοχές των βιοεπιστημών και της βιοιατρικής έχουν οδηγήσει σε έναν τεράστιο όγκο από δεδομένα στην περιοχή λογοτεχνίας, ο οποίος επεκτείνεται συνεχώς και στο μέγεθος και τη θεματική κάλυψη. Με το συντριπτικό ποσό κειμενικών πληροφοριών που παρουσιάζεται στην επιστημονική λογοτεχνία, υπάρχει μια ανάγκη για την αποτελεσματική αυτοματοποιημένη επεξεργασία που μπορεί να βοηθήσει τους επιστήμονες για να εντοπίσει, να συλλέξει και να χρησιμοποιήσει τη γνώση που κωδικοποιείται στην ηλεκτρονικά διαθέσιμη βιβλιογραφία . Το **MEDLINE** περιέχει πάνω από 14 εκατομμύρια αρχεία, που επεκτείνουν την κάλυψη του με περισσότερες από 40.000 περιλήψεις κάθε μήνα. Οι ανοικτοί εκδότες πρόσβασης όπως το **Biomed Central** έχουν αυξανόμενες συλλογές των ολοκληρωμένων κειμένων επιστημονικών άρθρων.

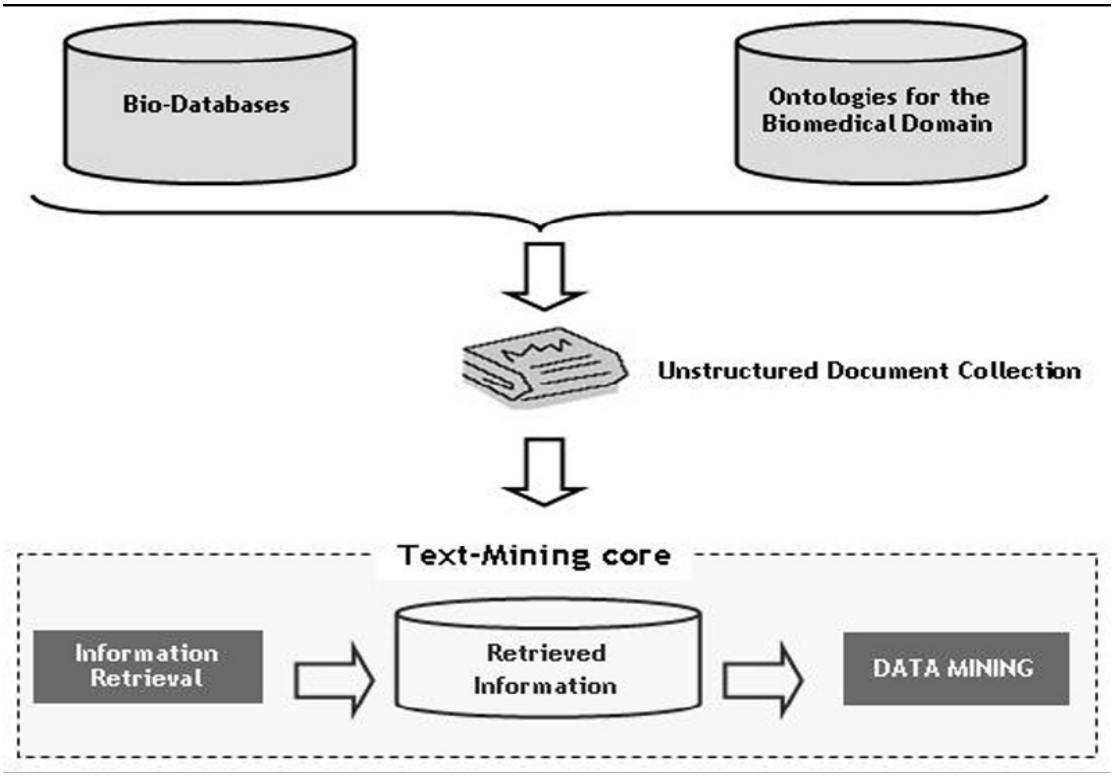
Επομένως, οι τεχνικές για τη εξόρυξη λογοτεχνίας δεν είναι πλέον μια επιλογή, αλλά μια προϋπόθεση για την αποτελεσματική ανακάλυψη γνώσης, διαχείριση, συντήρηση και αναπροσαρμογή μακροπρόθεσμα.
Για να επεξηγήσουν την αυξανόμενη κλίμακα του στόχου που αντιμετωπίζουν οι ειδικοί που προσπαθούν να ανακαλύψουν τις ακριβείς πληροφορίες ενδιαφέροντος μέσα στο biobibliome, θέτοντας μια ερώτηση όπως *η «θεραπεία καρκίνου του μαστού»* που υποβλήθηκε στη μηχανή αναζήτησης Medline το 2004 επέστρεψε σχεδόν 70.000 αναφορές ενώ οδήγησε σε 20.000 περιλήψεις το 2001. Η αποτελεσματική διαχείριση των βιοϊατρικών πληροφοριών είναι, επομένως, ένα κρίσιμο ζήτημα, έτσι οι ερευνητές πρέπει να είναι σε θέση να επεξεργαστούν τις πληροφορίες και γρήγορα και συστηματικά. Παραδοσιακά, οι βιοερευνητές ψάχνουν τη βιοϊατρική λογοτεχνία χρησιμοποιώντας τη διεπαφή PUBMED για να ανακτήσουν τα έγγραφα MEDLINE.

2.3 PUBMED

PUBMED είναι μια αποθήκη ευρετηρίασης και ανάκτησης που διαχειρίζεται αρκετά εκατομμύρια έγγραφα. Αυτά τα έγγραφα συντάσσονται με το χέρι, όπου οι όροι δεικτών επιλέγονται και ορίζονται στα έγγραφα από ένα τυποποιημένο ελεγχόμενο λεξιλόγιο. Η ανάκτηση εφαρμόζεται σαν αναζήτηση λέξης κλειδιού, και έτσι τα έγγραφα που ικανοποιούν πλήρως μια ερώτηση ανακτώνται. Ένα άλλο πρόβλημα είναι η επιλογή των κατάλληλων όρων δεικτών που θα ανακτούσαν τα πιο σχετικά έγγραφα. Οι όροι δεικτών δεν χαρακτηρίζουν απαραιτήτως τα έγγραφα σημασιολογικά, αλλά χρησιμοποιούνται για να κάνουν διακρίσεις μεταξύ των εγγράφων. Οι κλασικές τεχνικές ανάκτησης πληροφοριών (IR – information retrieval) δεν χρησιμοποιούν οποιεσδήποτε γλωσσικές τεχνικές για να αντιμετωπιστεί η γλωσσική μεταβλητότητα όπως η συνωνυμία και η πολυσημία που μπορούν να παραγάγουν πολλά ψεύτικα θετικά. Ακόμη και οι ελεγχόμενες προσεγγίσεις ευρετηρίασης είναι ασυμβίβαστες και περιορίστηκαν δεδομένου ότι οι αποθήκες γνώσης είναι στατικές και δεν μπορούν να αντιμετωπίσουν τη δυναμική φύση των εγγράφων.

2.4 Εξόρυξη Κειμένου – ΕΚ (ορισμός)

Ορίζεται ως η διαδικασία ανακάλυψης και εξαγωγής γνώσης από μη δομημένη πληροφορία, αντιπαραβαλλόμενη με την εξόρυξη δεδομένων (data mining), η οποία ανακαλύπτει γνώση από δομημένη πληροφορία (Hearst, 1999). Αντί να αφήνουμε το χρήστη με το πρόβλημα να έχει να διαβάσει διάφορες δεκάδες χιλιάδων εγγράφων, με την εξόρυξη κειμένου δίνεται η δυνατότητα της εξαγωγής ακριβών γεγονότων, και την εύρεση ενδιαφερόντων συσχετισμών μεταξύ των ανόμοιων γεγονότων, οδηγώντας έτσι στην ανακάλυψη νέας ή ανυποψίαστης και κρυμμένης γνώσης στις αναφορές κειμένων .
Κανονικά η ΕΚ περιλαμβάνει τρία στάδια (Σχήμα 2.4.1) :



Σχήμα 2.1: Μια πιθανή άποψη των τμημάτων και των διαδικασιών εξόρυξης Κειμένου [B.Mathiak, S. Eckstein,"Five Steps to Text Mining in Biomedical Literature", Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics, pp 47-50].

1. Σε πρώτη φάση συλλέγονται οι σχετικές κείμενο – αναφορές, κυρίως βασισμένες σε μεθόδους εξαγωγής πληροφοριών.
2. Στο επόμενο βήμα, γνωστό ως Εξαγωγή Πληροφοριών, εκτελείται προσδιορισμός και εξαγωγή πληροφοριών από τμήματα πληροφοριών (κυρίως όρους ή μικρές φράσεις), από τα ανακτημένα κείμενα. Αυτό γίνεται σύμφωνα με τα αιτήματα του χρήστη, βασίζεται σε αρχές Ανάκτησης Πληροφοριών και κυρίως στις διαδικασίες ανάλυσης κειμένων.
3. Στο τελευταίο βήμα υιοθετούνται κυρίως τεχνικές εξόρυξης δεδομένων, μηχανική μάθηση και στατιστικές τεχνικές, προκειμένου να προκληθούν και να προσδιοριστούν οι συσχετισμοί μεταξύ των κομματιών των εξαγόμενων πληροφοριών.

2.5 Εξόρυξη Κειμένου και Ανάκτηση Πληροφοριών

Η **Ανάκτηση Πληροφοριών** (ΑΠ) σχετίζεται με την εντόπιση πληροφοριών σύμφωνα με τις ανάγκες του χρήστη. Ο πρωταρχικός στόχος στην ΑΠ είναι η λειτουργία της ανάκτησης όπου διάφορα διαθέσιμα έγγραφα που αναφέρονται σε μια συγκεκριμένη περιοχή θέματος, χρησιμοποιούνται για την έρευνα συγκεκριμένων όρων (κάτι ανάλογο με ένα ερευνητή που κάμνει μια αναζήτηση λογοτεχνίας σε μια βιβλιοθήκη). Σε αυτό το περιβάλλον το σύστημα ανάκτησης γνωρίζει το σύνολο εγγράφων στο οποίο θα ψάξει, αλλά δε μπορεί να προβλέψει ποιο θα είναι το συγκεκριμένο θέμα το οποίο θα ερευνηθεί.
Οι τεχνικές ΑΠ παρουσιάζουν τη (βασική) υποδομή (προσεγγίσεις, τεχνικές, συστήματα και εργαλεία), για τους μηχανισμούς εξόρυξης κειμένων.

2.6 Εξόρυξη Κειμένου στη Βιοιατρική Περιοχή

Η έρευνα μας επιλέγει ως πεδίο δεδομένων τις βιοιατρικές βάσεις δεδομένων. Συνεπώς θα δούμε παρακάτω την ανάλυση και ολοκλήρωση διαφόρων μεθόδων εξόρυξης κειμένων σε βιοιατρική βιβλιογραφία.

Η εξόρυξη κειμένων από βιοιατρικές βάσεις δεδομένων έχει εξελιχθεί τα τελευταία χρόνια και αποτελεί ένα σημαντικό εργαλείο στη βιοπληροφορική. Χωρίζουμε αυτή τη διαδικασία της εξόρυξης κειμένων σε πέντε ξεχωριστά στάδια:

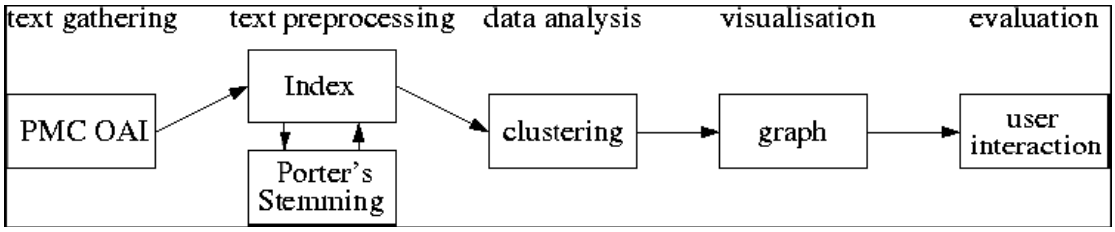
1. συλλογή-μάζεμα κειμένων
2. προ-επεξεργασία κειμένου
3. ανάλυση δεδομένων
4. οπτικοποίηση
5. αξιολόγηση

Κατά τη συλλογή κειμένων στο βιοιατρικό πεδίο ο χρήστης αρχικά λοιπόν δίνει μια πρώτη κατεύθυνση στο σύστημα σε ποια θεματική ενότητα να ψάξει για να βρει τα απαιτούμενα έγγραφα από τα οποία θα γίνει σε μετέπειτα στάδιο η εξόρυξη. Με αυτό τον τρόπο ο αριθμός των εγγράφων που θα ελεγχθούν ελαχιστοποιείται.

Το πρώτο στάδιο, της συλλογής κειμένων, επικεντρώνεται κυρίως στο μάζεμα των κειμένων που μας ενδιαφέρουν για την έρευνα μας σύμφωνα με το θεματικό πεδίο που έχουμε θέσει και τη μετατροπή των κειμένων στη κατάλληλη μορφή με κατάλληλες μεθόδους ούτως ώστε να διευκολυνθεί η εξαγωγή. Στο στάδιο της προ-επεξεργασίας γίνεται μια επεξεργασία του κειμένου κατά την οποία το κείμενο χωρίζεται σε λέξεις ή όρους(διακριτικά). Κατά αυτό τον τρόπο ξεχωρίζουμε τους όρους που μας ενδιαφέρουν περισσότερο. Έπειτα ακολουθούνται κάποιες μέθοδοι οι οποίες δίδουν την ακριβή σχέση μεταξύ των λέξεων. Ακόμα στόχος αυτού του σταδίου είναι να βελτιστοποιήσει την απόδοση του επόμενου σταδίου, της ανάλυσης δεδομένων.

Στην ανάλυση δεδομένων, δίδεται έμφαση στη γρήγορη συσταδοποίηση των κειμένων και στο μετέπειτα στάδιο της οπτικοποίησης, κυρίως γίνεται οπτική απεικόνιση των κειμενικών συστάδων που προκύπτουν από το στάδιο τρία και των αποτελεσμάτων που προκύπτουν με την εξόρυξη κειμένου γενικά.

Και τέλος στην αξιολόγηση, δίνουμε βάρος στις ειδικές δυνατότητες αξιολόγησης, όπως η αλληλεπίδραση με το χρήστη.



Σχήμα 2.2: Εφαρμογή παραδείγματος όπου γίνεται ανίχνευση θέματος συστάδας σε PMC OAI κείμενα πλήρης μορφής. [B.Mathiak, S. Eckstein,"Five Steps to Text Mining in Biomedical Literature", Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics, pp 47-50].

2.7 Εφαρμογές της Εξαγωγής Κειμένου στη Βιοιατρική

Υπάρχουν διάφορα βιοιατρικά παραδείγματα, όπου έχει εφαρμοστεί η εξαγωγή δεδομένων με επιτυχία. Παραδείγματα όπως η διάγνωση ασθενειών, όπου η εξαγωγή δεδομένων συσχετίζει τα συμπτώματα και άλλα χαρακτηριστικά των ασθενών σύμφωνα με την ασθένεια τους, υποομάδες από ασθενείς οι οποίοι είναι σε κίνδυνο για κάποια συγκεκριμένη ασθένεια, για την έκφραση κάποιου γονιδίου, με ένα αυξανόμενο αριθμό εφαρμογών, όπου οι προβλέψεις και οι προσδιορισμοί των δεικτών ασθενειών γίνονται, βασισμένοι στα χαρακτηριστικά γνωρίσματα των γονιδίων.

Ενώ όμως η εξαγωγή δεδομένων συνήθως λειτουργεί με τις συλλογές καλά δομημένων στοιχείων (structured data), οι ερευνητές πρέπει συχνά να εξετάσουν επίσης και ημι-δομημένες συλλογές κειμένων (semi – structured). Τέτοια σύνολα δεδομένων απαιτούν τη χρήση τεχνικών εξόρυξης κειμένων. Εξάγοντας σημαντικές πληροφορίες από την όλο και περισσότερο διαθέσιμη βιοιατρική γνώση που αναπαρίσταται με ψηφιακές μορφές κειμένων, αποδεικνύεται ως σημαντική ευκαιρία για βιοιατρικές ανακαλύψεις και παραγωγή υποθέσεων. Η Βιοιατρική Πληροφορική παρουσιάζεται κατά συνέπεια ως ένα απαραίτητο στοιχείο της βιοιατρικής ερευνητικής διαδικασίας.

Μέθοδοι που χρησιμοποιήθηκαν πρόσφατα για βιοιατρικούς στόχους εξαγωγής κειμένων, περιλαμβάνουν τα ακόλουθα στοιχεία:

- Named Entity Recognition, προκειμένου να προσδιοριστούν όλα τα στιγμιότυπα ενός ονόματος για ένα συγκεκριμένο τύπο περιοχής, μέσα σε μια συλλογή κειμένου.

Παραδείγματα πρόσφατων περιοχών βιοιατρικής έρευνας:

- ονόματα φαρμάκων εντός δημοσιευμένων άρθρων περιοδικού
- ονόματα γονιδίων και τα σύμβολα τους από συλλογή περιλήψεων του MEDLINE

Προσεγγίσεις Εξόρυξης Κειμένου:

- βασισμένες στις λέξεις, σε κανόνες, σε στατιστικές, συνδυασμός των πιο πάνω

- Κατηγοριοποίηση Κειμένων, με στόχο να αποφασίζει αυτόματα εάν ένα έγγραφο ή μέρος αυτού, έχει συγκεκριμένα χαρακτηριστικά ενδιαφέροντος.

Παραδείγματα πρόσφατων περιοχών βιοιατρικής έρευνας:

- Έγγραφα που συζητούν ένα δεδομένο θέμα
- Κείμενα που περιέχουν ένα συγκεκριμένο τύπο πληροφοριών

Προσεγγίσεις Εξόρυξης Κειμένου:

- ταξινόμηση με κανόνα επαγωγής

- Εξαγωγή συνωνύμου και συντμήσεων με στόχο την επιτάχυνση της αναζήτησης λογοτεχνίας με αυτόματες συλλογές συνωνύμων και συντμήσεων για οντότητες.

Παραδείγματα πρόσφατων περιοχών βιοιατρικής έρευνας:

- Συνώνυμα ονόματα γονιδίων
- Συντμήσεις βιοιατρικών όρων

Προσεγγίσεις Εξόρυξης Κειμένου:

- συνδυασμός ονομασμένων συστημάτων αναγνώρισης οντοτήτων, με στατιστικούς, SVM κατηγοριοποιητή, και αυτόματο ή χειρονακτικό βασισμένο σε πρότυπα, αλγόριθμο ταιριάσματος κανόνων.

- Εξαγωγή Σχέσης, με στόχο την αναγνώριση περιστατικών ενός προ-διευκρινισμένου τύπου μιας σχέσης μεταξύ ενός ζευγαριού οντοτήτων συγκεκριμένων τύπων.

Παραδείγματα πρόσφατων περιοχών βιοιατρικής έρευνας:

- Σχέσεις μεταξύ γονιδίων και πρωτεϊνών
- Συσταδοποίηση γονιδίων βασισμένη σε κείμενο

Προσεγγίσεις Εξόρυξης Κειμένου:

- ανάλυση απόκλισης γείτονα, προσέγγιση με διανυσματικά διαστήματα και k-medoids αλγόριθμο συσταδοποίησης, συγκεχυμένη θεωρία συνόλων πάνω σε καθορισμένα αρχεία δεδομένων

- Πλαίσια Ολοκλήρωσης, με στόχο να εξετάσει πολλές διαφορετικές ανάγκες των χρηστών.

Παραδείγματα πρόσφατων περιοχών βιοιατρικής έρευνας:

- Σύγκριση των ονομάτων των γονιδίων και λειτουργικών όρων

- Γονίδια βασισμένο σε συστάδες κειμένου

Προσεγγίσεις Εξόρυξης Κειμένου:

- Βασισμένες σε πρότυπα, βασισμένες σε συστάδες και text-profiling

- Παραγωγή Υπόθεσης, η οποία εστιάζεται στην αποκάλυψη των υπονοούμενων σχέσεων, αντάξια περεταίρω έρευνας, το οποίο προκύπτει από την παρουσία άλλων πιο ρητών πληροφοριών.

Παραδείγματα πρόσφατων περιοχών βιοιατρικής έρευνας:

- Πιθανές νέες χρήσεις και θεραπευτικά αποτελέσματα από τα φάρμακα

Προσεγγίσεις Εξόρυξης Κειμένου:

- Βασισμένες στο Swanson's ABC μοντέλο

2.8 Μηχανική Μάθηση

Η μηχανική μάθηση ασχολείται με το σχεδιασμό και την ανάπτυξη αλγορίθμων και τεχνικών που να επιτρέπουν στον υπολογιστή «να μαθαίνει». Γενικά υπάρχουν δύο τύποι μαθήσεων, ο επαγωγικός και ο παραγωγικός τύπος.

Η σημαντικότερη εστίαση της έρευνας μηχανικής μάθησης είναι η εξαγωγή πληροφοριών από δεδομένα αυτόματα, με υπολογιστικές και στατιστικές μεθόδους. Ως εκ τούτου η μηχανική μάθηση συσχετίζεται πολύ, όχι μόνο με την εξαγωγή δεδομένων και τις στατιστικές, αλλά και τη θεωρητική πληροφορική.

Η μηχανική μάθηση έχει ένα ευρύ φάσμα εφαρμογών, έτσι θα τη δούμε και εδώ στη συγκεκριμένη μελέτη να βρίσκει εφαρμογή.

2.9 Συμπεράσματα

Στο κεφάλαιο αυτό μελετήσαμε κάποιες από τις βασικότερες έννοιες και ορολογίες που αποτέλεσαν τη αρχή της έρευνας και μελέτης μου για την κατανόηση ενός πιο εξειδικευμένου θέματος που θα μας απασχολήσει στη συνέχεια, το οποίο στηρίζεται σε όλα αυτά που είδαμε εδώ. Και αυτό το ενδιαφέρον θέμα το οποίο θα δούμε, θα μελετήσουμε, θα αναπτύξουμε και θα το δούμε να υλοποιείται μέσα από κάποιο σύστημα είναι η Οντολογίες και συγκεκριμένα η Κατασκευή Οντολογιών.

Κεφάλαιο 3

ΟΝΤΟΛΟΓΙΕΣ &

ΔΗΜΙΟΥΡΓΙΑ ΟΝΤΟΛΟΓΙΩΝ

3.1 Εισαγωγή

Στο κεφάλαιο αυτό θα επιχειρήσουμε να δώσουμε κάποιους πρώτους ορισμούς για το τι είναι οντολογία καθώς και διάφορους τύπους οντολογιών. Επίσης δίδουμε κάποια εισαγωγικά περί της δημιουργίας οντολογιών και τη σχέση που έχει η οντολογία με την ανακάλυψη γνώσης καθώς και μια σύντομη αναφορά στο σύστημα OntoGen. Ένα σύστημα δημιουργίας οντολογιών. Ακόμα γίνεται μια αναφορά στη βιβλιοθήκη λογισμικού εξόρυξης κειμένων Text Garden.

3.2 Οντολογία

Οντολογία

Ένας ορισμός που δόθηκε από την αρχαιότητα για την οντολογία, ως

« η μεταφυσική μελέτη της φύσης της ζωής και της ύπαρξης»

Και μετέπειτα η έννοια της οντολογίας υιοθετείται από την Τεχνητή Νοημοσύνη και σημαίνει ειδικότερα

«μια διαμοιρασμένη και κοινή κατανόηση κάποιου τομέα, η οποία μπορεί να ανταλλαχθεί μεταξύ ανθρώπων και συστημάτων εφαρμογών» (Gruber).

Στο κλάδο λοιπόν της Πληροφορικής η οντολογία είναι μια αντιπροσώπευση ενός συνόλου εννοιών που εμπίπτουν σε μια περιοχή και σχέσεων μεταξύ αυτών των εννοιών. Χρησιμοποιείται έτσι ώστε να δικαιολογήσει τις ιδιότητες εκείνης της περιοχής, και μπορεί να χρησιμοποιηθεί για να καθορίσει την περιοχή εκείνη.

Οι οντολογίες χρησιμοποιούνται στην Τεχνητή Νοημοσύνη, στο Σημασιολογικό Ιστό, στη Τεχνολογία Λογισμικού, στην Βιοιατρική Ιατρική όπου και θα τη δούμε, στην Επιστήμη Βιβλιοθηκών και στην Αρχιτεκτονική Πληροφοριών ως μορφή αναπαράστασης της γνώσης για τον κόσμο ή για κάποιο μέρος αυτού.

Τα κοινά συστατικά που συναντούμε σε όλες τις οντολογίες είναι:

- **Κατηγορίες**: ιδιότητες, χαρακτηριστικά γνωρίσματα, χαρακτηριστικά ή παράμετροι που μπορεί να έχουν τα αντικείμενα (και οι κατηγορίες).

- **Σχέσεις**: οι τρόποι με τους οποίους οι κατηγορίες και τα αντικείμενα μπορούν να συσχετίζονται το ένα με το άλλο.

- **Όροι Λειτουργίας**: σύνθετες δομές που διαμορφώνονται από ορισμένες σχέσεις που μπορούν να χρησιμοποιηθούν αντί ενός μεμονωμένου όρου σε μια δήλωση.

- **Περιορισμοί**: τυπικά δηλωμένες περιγραφές για το τι πρέπει να είναι αληθής για ένα ισχυρισμό, έτσι ώστε να γίνεται αποδεκτός ως είσοδος.

- **Κανόνες**: δηλώσεις της μορφής if-then πρότασης, που περιγράφουν τα λογικά συμπεράσματα που μπορούν να προέλθουν από ένα ισχυρισμό σε μια συγκεκριμένη μορφή.

- **Axia**: ισχυρισμοί, συμπεριλαμβανομένων των κανόνων, σε μια λογική μορφή μαζί με τη γενική θεωρία που περιγράφει η οντολογία, για την περιοχή εφαρμογής της.

- **Γεγονότα:** η αλλαγή των ιδιοτήτων ή των σχέσεων.

Οι Οντολογίες κωδικοποιούνται συνήθως χρησιμοποιώντας **γλώσσες οντολογίας**. Μια Γλώσσα Οντολογίας είναι μια επίσημη γλώσσα (formal language) που χρησιμοποιείται για να κωδικοποιήσει την οντολογία. Υπάρχουν διάφορες τέτοιες γλώσσες για τις οντολογίες, όπως οι OWL, KIF, CycL, Gellish και άλλες.

3.3 Δημιουργία Οντολογίας

Όταν αναφερόμαστε σε οντολογίες, συνήθως αναφερόμαστε σε μια οντολογία σαν αναπαράσταση ενός γράφου/δομή δικτύου, που αποτελείται από:

1. ένα σύνολο από έννοιες (στο γράφο εμφανίζονται με τη μορφή vertices)
2. ένα σύνολο από σχέσεις οι οποίες ενώνουν τις έννοιες (στο γράφο εμφανίζονται με απευθείας ακμές)
3. ένα σύνολο από στιγμιότυπα, τα οποία ανατίθενται σε κάποιες συγκεκριμένες έννοιες

Ένας πιο ακριβής ορισμός για την οντολογία εξαρτάται από το περιεχόμενο στο οποίο χρησιμοποιείται η οντολογία. Τα στιγμιότυπα, οι έννοιες και οι σχέσεις μπορεί να παρουσιάζονται με πολύ διαφορετική μορφή σε διαφορετικές οντολογίες.

Παραδείγματα μερικών από τους πιο πολύ ευρέως χρησιμοποιούμενους τύπους οντολογιών, είναι:

- Οντολογίες Ορολογιών, όπου οι έννοιες λέξεις αισθήσεων και τα στιγμιότυπα είναι λέξεις (όπως η οντολογία Word Net).
- Θεματικές Οντολογίες, όπου οι έννοιες είναι θέματα και τα στιγμιότυπα είναι έγγραφα (Όπως η DMoz. Με δημιουργία θεματικών οντολογιών θα ασχοληθούμε στη συνέχεια της διπλωματικής εργασίας).
- Οντολογία Μοντέλου Δεδομένων, όπου οι έννοιες είναι πίνακες σε μια βάση δεδομένων και τα στιγμιότυπα εγγραφές δεδομένων (όπως σε μια βάση δεδομένων).

Γενικά οι οντολογίες μπορούν να είναι πιο περίπλοκα αντικείμενα από αυτά που περιγράφηκαν πιο πάνω – οι έννοιες και οι σχέσεις για παράδειγμα μπορούν να περιγραφούν σε first-order logic, τα στιγμιότυπα μπορούν να είναι αυθαίρετα σύνθετα αντικείμενα.

Από την προοπτική της χρησιμοποίησης των μεθόδων της Ανακάλυψης Γνώσεων για την αυτόματη ή ημι – αυτόματη δημιουργία οντολογιών, το σημαντικό μέρος είναι ο προσδιορισμός του τρόπου με τον οποίο γίνεται η σύγκριση των στιγμιότυπων μεταξύ τους, ούτως ώστε να διαμορφώσει τις έννοιες και πώς να βάλει τις σχέσεις μεταξύ τους. Στην Ανάκτηση Πληροφοριών και Εξαγωγή Κειμένων, έχουμε διάφορα μέτρα για εκτίμηση της ομοιότητας μεταξύ των εγγράφων, καθώς επίσης και για την ομοιότητα μεταξύ των αντικειμένων που χρησιμοποιούνται μέσα στα έγγραφα.

Ο λειτουργικός καθορισμός μιας οντολογίας από την προοπτική της Ανακάλυψης Γνώσεων πρέπει να είναι μάλλον τεχνικός και συνοπτικός στο βαθμό που επιτρέπει τον υπολογισμό της πολυπλοκότητας των διαφόρων ημι – αυτόματων προσεγγίσεων κατασκευών οντολογίας. Στην ορολογία μηχανική μάθηση, αναφερόμαστε στην

δημιουργία οντοτήτων ως μάθηση οντολογίας. Συνεπώς, ορίζουμε μια οντολογία ακριβώς ως μια άλλη κατηγορία μοντέλων (ελαφρώς πιο σύνθετη συγκρινόμενη με τα συνηθισμένα μοντέλα Μηχανικής Μάθησης), η οποία πρέπει να εκφραστεί σε κάποιου είδους γλώσσα υπόθεσης. Αυτός ο ορισμός της μάθησης οντολογίας περιλαμβάνει τα ακόλουθα υπό – προβλήματα που είναι σχετικά στα διαφορετικά πλαίσια:

1. μάθηση ακριβώς μόνο των εννοιών
2. μάθηση ακριβώς μόνο των σχέσεων μεταξύ των υπαρχουσών εννοιών
3. μάθηση των εννοιών και των σχέσεων, συγχρόνως
4. επέκταση μιας υπάρχουσας οντολογίας/δομής
5. αντιμετώπιση με δυναμικά ρεύματα δεδομένων

Η γλώσσα η οποία χρησιμοποιούμε για να εκφράσουμε τις έννοιες και τις σχέσεις τους, καθορίζει την πολυπλοκότητα και τη δύναμη της διαδικασίας μάθησης. Σε ευρύτερο πλαίσιο, η προσέγγιση Ανακάλυψης Γνώσεων στην αυτόματη ή ημι – αυτόματη δημιουργία οντολογίας εξετάζει κάποια είδη αντικειμένων από δεδομένα τα οποία πρέπει να έχουν κάποιες ιδιότητες – μπορεί να είναι έγγραφα, εικόνες, αρχεία δεδομένων ή συνδυασμός των πιο πάνω.

Έχουν χρησιμοποιηθεί διάφορες προσεγγίσεις για το κτίσιμο οντολογιών, οι περισσότερες από τις οποίες χρησιμοποιούν κυρίως χειροκίνητες μεθόδους.

Μια προσέγγιση στην οικοδόμηση οντολογιών τέθηκε στη μελέτη CYC (Lenat and Guha, 1990), όπου το κύριο βήμα εμπεριέχει την χειρωνακτική εξαγωγή της γνώσης από τις διάφορες πηγές. Έχουν υπάρξει μερικοί ορισμοί της μεθοδολογίας για το κτίσιμο μιας οντολογίας, υποθέτοντας πάλι τη χειρονακτική προσέγγιση. Για παράδειγμα, η μεθοδολογία που προτάθηκε (Ushold and King, 1995) περιλαμβάνει τα ακόλουθα στάδια: προσδιορισμός του σκοπού της οντολογίας (γιατί να κτιστεί, πως θα χρησιμοποιηθεί, φάσμα των χρηστών), κτίσιμο της οντολογίας, αξιολόγηση και τεκμηρίωση.

Το κτίσιμο της οντολογίας, διαιρείται περαιτέρω σε τρία βήματα. Το πρώτο είναι η σύλληψη της οντολογίας, στο οποίο προσδιορίζονται οι βασικές έννοιες, ένας ακριβής ορισμός αυτών γράφεται σε κειμενική μορφή. Επίσης προσδιορίζονται οι όροι που θα χρησιμοποιηθούν για να αναφερθούν στις έννοιες και στις σχέσεις. Το δεύτερο βήμα, περιλαμβάνει την κωδικοποίηση της οντολογίας για να αναπαραστήσει την ορισμένη εννοιολογία σε κάποια επίσημη γλώσσα (formal language). Το τρίτο στάδιο, περιλαμβάνει, την πιθανή ολοκλήρωση με τις υπάρχουσες οντολογίες. Μια επισκόπηση των μεθοδολογιών για το κτίσιμο των οντολογιών παρέχεται μέσα από (Fernandez, 1999), όπου διάφορες μεθοδολογίες, συμπεριλαμβανομένης αυτής που περιγράφηκε πιο πάνω, παρουσιάζονται και αναλύονται ενάντια στις IEEE Standard for Developing Software Life Cycle Processes, όπου εξετάζει τις οντολογίες ως μέρη κάποιου προϊόντος λογισμικού.

Ειδικότερα θα δούμε σε παρακάτω κεφάλαια το σύστημα OntoGen, το οποίο κατασκευάζει οντολογίες και χαρακτηρίζεται ως «ημι-αυτόματο (semi-automatic)» και «data-driven», δηλαδή με τον όρο «ημι-αυτόματο», εννοούμε ότι το σύστημα είναι εργαλείο που αλληλεπιδρά με το χρήστη ούτως ώστε να τον βοηθά κατά τη διαδικασία της δημιουργίας της οντολογίας. Το σύστημα προτείνει τις έννοιες, τις σχέσεις και τα ονόματα τους, αναθέτει αυτόματα στιγμιότυπα στις έννοιες και παρέχει μια καλή επισκόπηση της οντολογίας. Συγχρόνως ο χρήστης μπορεί να ρυθμίσει πλήρως όλες τις ιδιότητες της οντολογίας χειροκίνητα προσθέτοντας ή αφαιρώντας έννοιες, σχέσεις και επαναθέτοντας στιγμιότυπα. Ο χαρακτηρισμός «data-driven» δίδεται στο σύστημα διότι το μεγαλύτερο μέρος της ενίσχυσης που παρέχεται από το σύστημα (έννοια, εισήγηση σχέσης, κλπ) είναι βασισμένο σε μερικά ελλοχεύοντα δεδομένα που παρέχονται από το χρήστη στην αρχή της δημιουργίας της οντολογίας. Τα δεδομένα απεικονίζουν την περιοχή (θεματική) για την οποία κτίζει ο χρήστης την οντολογία. Τα στιγμιότυπα εξάγονται από τα δεδομένα.

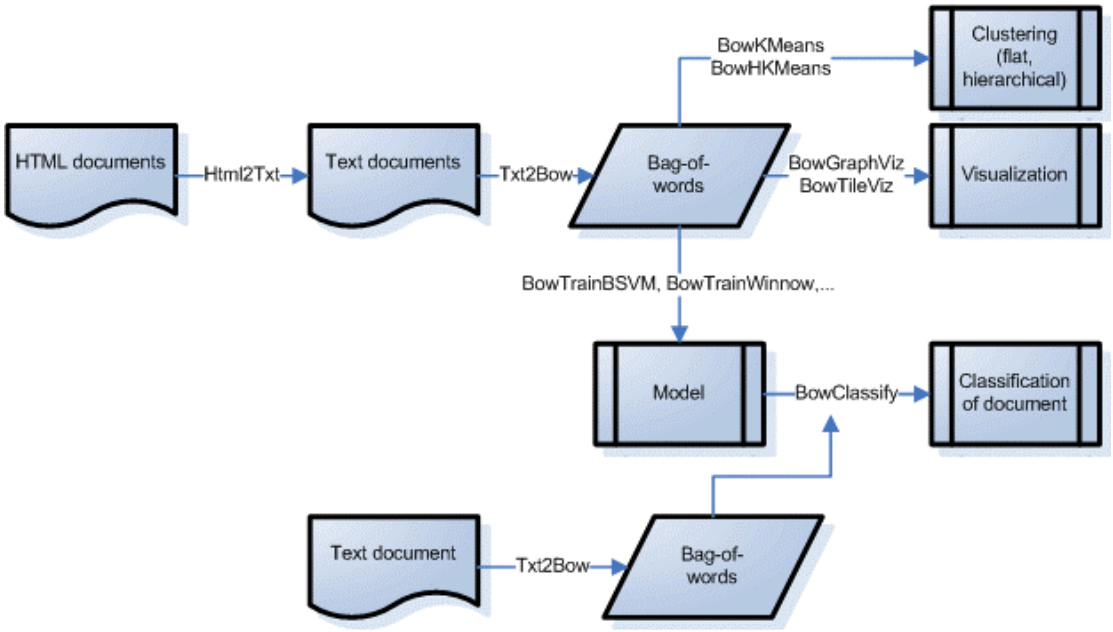
3.4 Βιβλιοθήκη Text Garden

Το σύστημα OntoGen και ειδικότερα η λειτουργικότητα του βασίζεται στη βιβλιοθήκη λογισμικού εξόρυξης κειμένου Text Garden και αναπτύχθηκε κάτω από τη μελέτη SEKT.

Η Text Garden είναι όπως ειπώθηκε και πιο πάνω μια βιβλιοθήκη λογισμικού με τα ακόλουθα τεχνικά χαρακτηριστικά:

- Η βιβλιοθήκη είναι γραμμένη σε C++ και απαιτεί Microsoft .NET framework 2.0 για να τρέξει.
- Η βιβλιοθήκη μεταγλωττίζει σε MS Visual C++ και Borland C++ Builder 3.0, 5.0 χωρίς οποιοδήποτε warning. Μελλοντικά σχέδια είναι να συνδέσουν τη βιβλιοθήκη στο εγγύς μέλλον και με Linux.
- Περιέχει πάνω από 100K γραμμές κώδικα που περιέχεται σε περίπου 200.h/.cpp αρχεία.

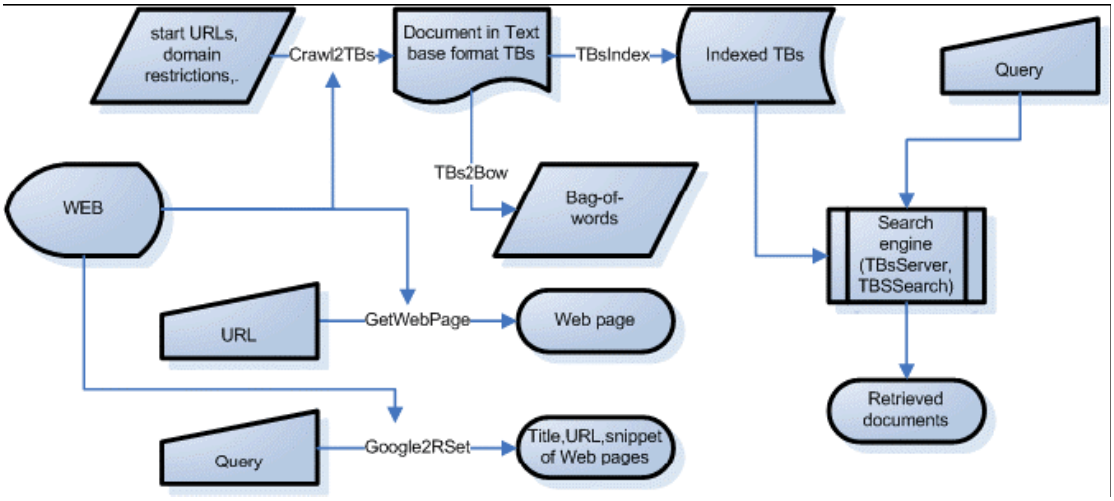
Η λειτουργία της βιβλιοθήκης Text Garden είναι χωρισμένη σε διάφορες ομάδες ανάλογα με το πώς χρησιμοποιείται κάθε φορά ως εξής. Προετοιμασία δεδομένων συμπεριλαμβανομένων μετατροπών, σύρσιμο(crawling) και άλλα παρόμοια. Μοντελοποίηση των δεδομένων συμπεριλαμβανομένου διαφορετικών αλγορίθμων για επιβλεπόμενη (Sebastian, 2002) και ημι – επιβλεπόμενη (Nigam et al, 2001) μάθηση (π.χ. SVM, k-means clustering). Αποδοτικές υλοποιήσεις πράξεων γραμμικής άλγεβρας που είναι σε θέση να αντιμετωπίσουν μεγάλα σύνολα δεδομένων. Απεικόνιση των δεδομένων (Σχήμα 3.1), αναζήτηση κειμενικών εγγραφών (Σχήμα 3.2). Το κύριο μέρος της λειτουργίας είναι το σύνολο των κατηγοριών εξόρυξης κειμένου που καλύπτουν διάφορες σωληνώσεις (Σχήματα 3.1,3.2,3.3). Βασικά, η ιδέα είναι να καλυφθεί το μεγαλύτερο μέρος της εξόρυξης κειμένου, της ανάκτησης πληροφοριών, του web-mining (Σχήμα 3.3), crawling/search σενάριο (Σχήμα 3.2) που μπορεί κάποιος να χρειαστεί σε πρακτικές εφαρμογές σε βασικές τεχνικές εξόρυξης κειμένου.



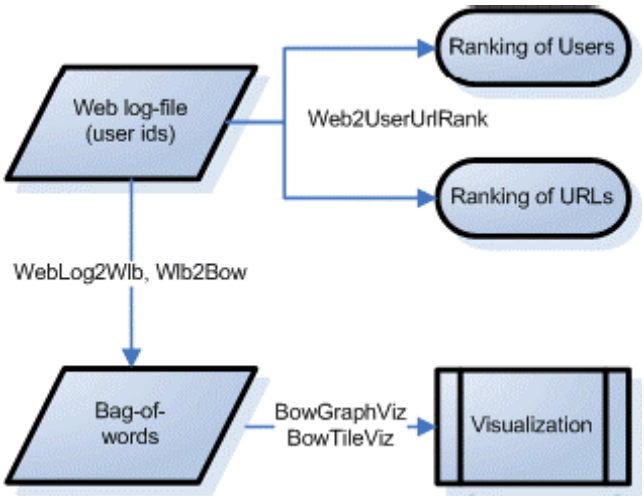
Σχήμα 3.1: Το κύριο σενάριο με τα αντίστοιχα συστατικά της βιβλιοθήκης Text Garden για την ταξινόμηση, τη συσταδοποίηση και την απεικόνιση των κειμενικών εγγράφων. [D. Mladenic, M. Grobelnik, "Knowledge Discovery for Ontology Discovery"].

3.4.1 Μορφή αρχείων που χρησιμοποιείται για την αναπαράσταση των εγγράφων

Υπάρχουν διάφορες μορφές για αναπαράσταση των κειμενικών εγγράφων που χρησιμοποιούνται από το Text Garden, καλύπτοντας έτσι πολλούς διαφορετικούς τρόπους διαχειρισμού κειμενικών εγγράφων. Μια μορφή που υποστηρίζει το σύστημα OntoGen είναι η Bag-of-Words format. Η μορφή αυτή συμπεριλαμβάνει τα έγγραφα σε μια πλήρη επεξεργασμένη μορφή όπου υποστηρίζεται αυτή η συγκεκριμένη μορφή. Κάθε έγγραφο αναπαριστάται από ένα σύνολο με τις συχνότητες των λέξεων και των κατηγοριών οι οποίες ανήκουν σε αυτό. Ο σκοπός του έγγραφου είναι να εκτελέσει μια αποδοτική εκτέλεση των αλγόριθμων σε συνδυασμό με την αναπαράσταση bag-of-words, όπως: κατηγοριοποίηση, μάθηση, οπτική απεικόνιση κ.λπ. Υπάρχει μια χρησιμότητα στο Text Garden (Txt2Bow) την οποία χρησιμοποιούμε για μετατροπή της μορφής του κειμένου στη μορφή Bag-Of-Words (".Bow").



Σχήμα 3.2: Το κύριο σενάριο με τα αντίστοιχα συστατικά της βιβλιοθήκης Text Garden για crawling στο Web [D. Mladenic, M. Grobelnik, "Knowledge Discovery for Ontology Discovery"].



Σχήμα 3.3: Το κύριο σενάριο με τα αντίστοιχα συστατικά της βιβλιοθήκης Text Garden για εξόρυξη των Web log files [D. Mladenic, M. Grobelnik, "Knowledge Discovery for Ontology Discovery"].

3.5 Συμπέρασμα

Έχουμε δει στο κεφάλαιο αυτό κάποιους πολύ εισαγωγικούς ορισμούς. Αυτό που μας ενδιαφέρει σε αυτή την εργασία είναι να πετύχουμε να ανακαλύψουμε γνώση μέσα από αυτή τη μεγάλη μάζα δεδομένων που μας κατακλύζει καθημερινά, αλλά με ένα τρόπο γρήγορο και συνεπώς έξυπνο. Με ένα τρόπο, με ένα σύστημα που θα το κάμνει αυτό, αντί για εμάς. Προτού δούμε αυτό το εργαλείο πρέπει να μελετήσουμε στα επόμενα κεφάλαια κάποιες τεχνικές που είναι απαραίτητες να διαθέτει το εργαλείο αυτό για να είναι ικανό να επεξεργάζεται αυτό το μεγάλο όγκο δεδομένων.

Κεφάλαιο 4

ΑΝΑΛΥΣΗ ΣΥΣΤΗΜΑΤΟΣ ONTOGEN :

Ημι – Αυτόματος Συντάκτης Οντολογιών

4.1 Εισαγωγή

Στο παρακάτω κεφάλαιο θα παρουσιάσω και θα αναλύσω τον συντάκτη οντολογιών όπως υλοποιείται στη τελευταία έκδοση του συστήματος OntoGen. Στο σύστημα θα δούμε να ενσωματώνονται μηχανική μάθηση και αλγόριθμοι εξόρυξης κειμένου γεγονός που οδηγεί σε ένα αποδοτικό περιβάλλον διεπαφής με το χρήστη, μειώνοντας συνεπώς το εμπόδιο εισόδου στο σύστημα από χρήστες οι οποίοι δεν είναι επαγγελματίες μηχανικοί οντολογίας. Τα κύρια χαρακτηριστικά του συστήματος

συμπεριλαμβάνουν επιβλεπόμενες και μη-επιβλεπόμενες μεθόδους για εισήγηση εννοιών και ονοματολογία εννοιών, όπως επίσης οπτικοποίηση οντολογιών και εννοιών. Το σύστημα έχει ελεγχθεί σε εκτενή δοκιμές χρηστών, καθώς και σε πάρα πολλά πραγματικού χρόνου σενάρια με πολύ θετικά αποτελέσματα.

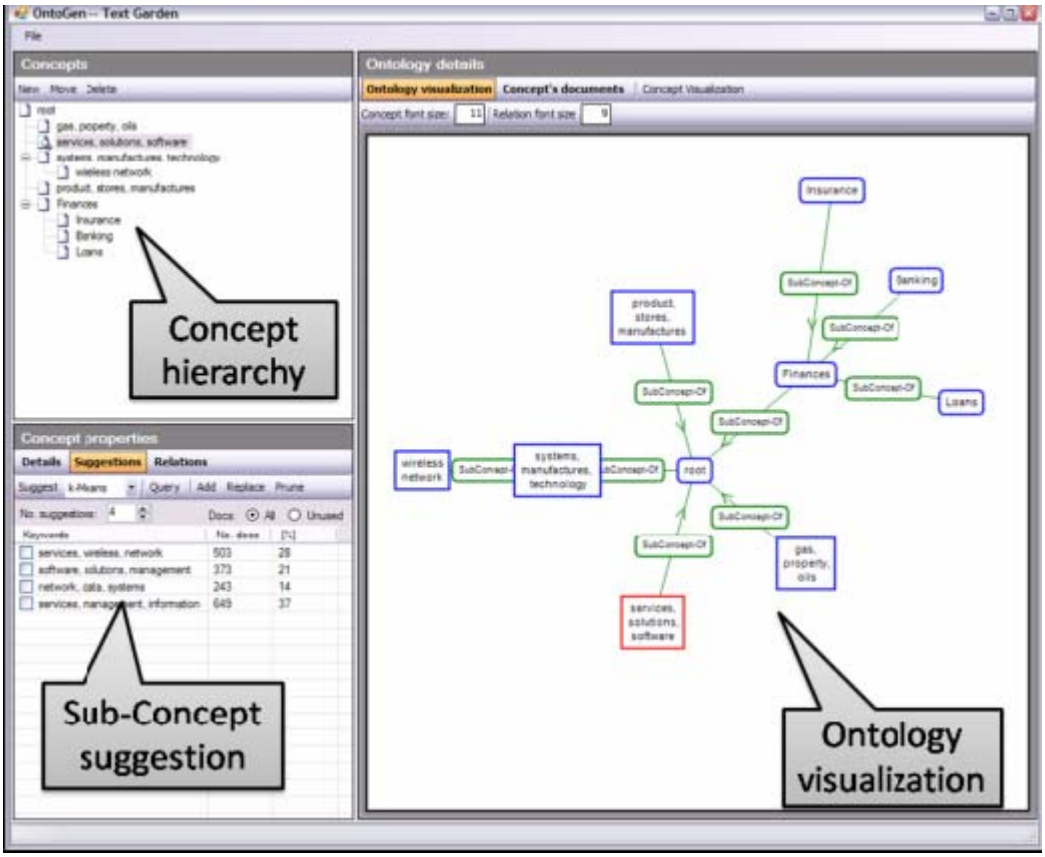
4.2 Το σύστημα OntoGen

Τα κύρια χαρακτηριστικά του συστήματος γενικά εξυπηρετούν τον ένα ή και τους δύο από τους κύριους στόχους του OntoGen :

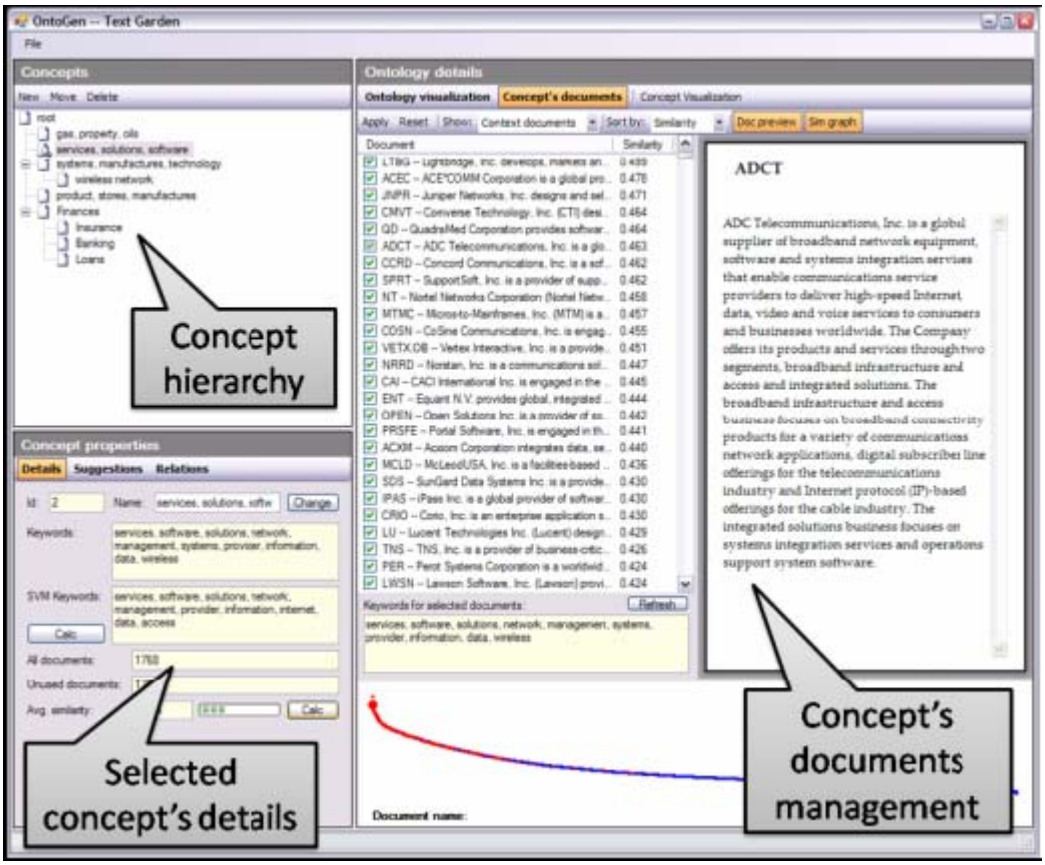
(1) απεικόνιση και ανίχνευση των ήδη υπαρχόντων εννοιών στην οντολογία
(2) πρόσθεση νέων εννοιών ή τροποποιήσεις στις ήδη υπάρχουσες έννοιες διαμέσου απλών καθοδηγητικών διαδικασιών υποβοηθούμενων από μηχανική μάθηση και τεχνικές εξόρυξης κειμένου.

Το κύριο παράθυρο του συστήματος παρέχει πολλαπλές όψεις της οντολογίας. Ένας φυσικός τρόπος αναπαράστασης της εννοιολογικής ιεραρχίας, που είναι κάπως πιο διαισθητικός προς τους περισσότερους χρήστες, είναι η αναπαράσταση με δέντρο. Αυτή η αναπαράσταση, είναι κάτι αντίστοιχο με αυτό που υπάρχει στο λειτουργικό σύστημα των Windows, όπου χρησιμοποιείται για να δείξει τη δομή των αρχείων και σαν μια απεικόνιση που προσφέρει μια άποψη της όλης οντολογίας με μια μόνο ματιά.
Κάθε έννοια από την οντολογία, εκτίθεται περαιτέρω (εννοώντας ότι περισσότερες λεπτομέρειες κάτω από αυτή γίνονται γνωστές), από το σύνολο των πιο πληροφοριακών λέξεων-κλειδιών για την έννοια στόχο (target concept), η οποία εξάγεται αυτόματα χρησιμοποιώντας τεχνικές εξόρυξης κειμένου και διαμέσου ενός θεματικού χάρτη(topic-map) των εγγράφων που ανήκουν σε αυτή την έννοια (περισσότερες λεπτομέρειες θα δούμε παρακάτω).

Στο Σχήματα 4.1 α και β έχουμε το τυποποιημένο σχεδιάγραμμα του συστήματος.



Σχήμα 4.1α : Ο χρήστης παίρνει εισηγήσεις για τις υπό-έννοιες για τις επιλεγμένες έννοιες(αριστερό κάτω μέρος), η οντολογία απεικονίζεται σε μια ιεραρχία σε μορφή κειμένου (αριστερό πάνω μέρος) και με γραφική μορφή (δεξί κεντρικό μέρος).



Σχήμα 4.1β: Ο χρήστης κάμνει μια έρευνα για την έννοια την οποία επέλεξε, κάνοντας browsing στις στατιστικές της έννοιας (κάτω αριστερά), τα έγγραφα που έχουν σχέση με την έννοια, καθώς και το περιεχόμενο τους εμφανίζονται και αυτά (πάνω αριστερά) και η γραφική που δείχνει τις ομοιότητες των εγγράφων για τη συγκεκριμένη έννοια(κάτω δεξιά).

Οι εισηγήσεις των εννοιών διαδραματίζουν ένα κύριο μέρος για το σύστημα. Για την παραγωγή εισηγήσεων, παρέχονται δύο μέθοδοι, η μη-επιβλεπόμενη και η επιβλεπόμενη μέθοδος.

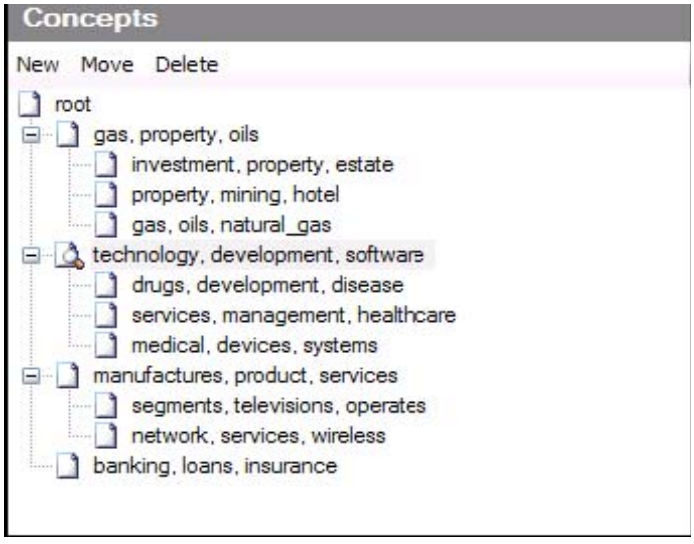
Οι μη-επιβλεπόμενοι μέθοδοι μάθησης, παράγουν αυτόματα μια λίστα από υπό-έννοιες για μια έννοια που επιλέγεται για εκείνη τη χρονική στιγμή, χρησιμοποιώντας τεχνικές k-means συσταδοποίησης και Latent semantic indexing (LSI), για να παράξουν μια λίστα από πιθανές υπό- έννοιες.

Οι επιβλεπόμενοι μέθοδοι μάθησης από την άλλη πλευρά, απαιτούν από το χρήστη να έχει μια πρόχειρη ιδέα σχετικά με ένα νέο θέμα – αυτό ορίζεται διαμέσου ενός query, μιας ερώτησης που επιστρέφουν τα έγγραφα. Το σύστημα ορίζει αυτόματα τα

έγγραφα που ανταποκρίνονται στο θέμα και η επιλογή μπορεί μετέπειτα να τελειοποιηθεί διαμέσου του χρήστη, αλληλεπιδρώντας κατάλληλα με το σύστημα, διαμέσου ενός ενεργού βρόγχου μάθησης χρησιμοποιώντας μια τεχνική μάθησης μηχανής για ημι – αυτόματη απόκτηση της γνώσεως του χρήστη. Επίσης το σύστημα παρέχει τρόπους για καθορισμό έννοιας, επιλέγοντας περιοχές στον ανάλογο θεματικό χάρτη.

4.3 Υλοποίηση των Κύριων Χαρακτηριστικών

Σε αυτό το τμήμα πρόκειται να περιγραφούν τα κύρια χαρακτηριστικά του OntoGen, συμπεριλαμβανομένου της απεικόνισης της ιεραρχίας των εννοιών, τη διαχείριση μιας έννοιας παρέχοντας τις λεπτομέρειες της έννοιας και τις εισηγήσεις για την ονομασία της, σχηματισμός της νέας έννοιας βασισμένος σε μη – επιβλεπόμενους και σε επιβλεπόμενους μεθόδους μηχανικής μάθησης καθώς και απεικόνιση εννοιών. Επίσης παρέχεται μια σύντομη περιγραφή κάποιων επιπρόσθετων χαρακτηριστικών τα οποία επιτρέπουν το collaborative τύπωμα των οντολογιών και την σκιαγράφηση των χρηστών (user profiling).



Σχήμα 4.2: Αναπαράσταση ιεραρχίας εννοιών με δενδροειδή μορφή. Ο χρήστης έχει τη δυνατότητα να επανατοποθετήσει μια ήδη υπάρχουσα έννοια σε διαφορετική θέση στην ιεραρχία ή να διαγράψει μια υπάρχουσα έννοια από την οντολογία. Όταν

μετακινείται μια έννοια, το πρόγραμμα προβάλλει ένα παράθυρο διαλόγου ζητώντας από το χρήστη να επιλέξει την νέα θέση για έννοια που πρόκειται να μετακινήσει.

4.3.1 Απεικόνιση της Ιεραρχίας Εννοιών

Το μέρος του κυρίως παράθυρου, το οποίο είναι πάντοτε ορατό στο χρήστη, είναι η απεικόνιση με δέντρο της εννοιολογικής ιεραρχίας (πάνω αριστερά στο Σχήμα 4.1 «Ιεραρχία Εννοιών», που απομονώθηκε στο Σχήμα 4.2). Προσφέρει μια γρήγορη άποψη όλων των εννοιών με τη θέση τους στην ιεραρχία των εννοιών, με άμεση πρόσβαση σε εντολές προς διαχειρισμό της ιεραρχίας.

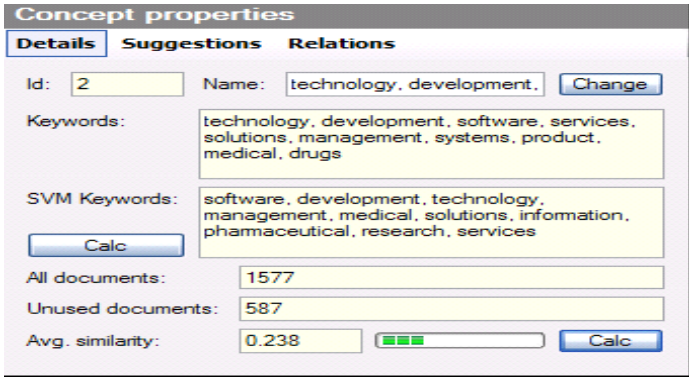
Επίσης η Οντολογία απεικονίζεται σε πραγματικό χρόνο στη δεξιά πλευρά του κυρίως παραθύρου («Ontology visualization» στο Σχήμα 4.1). Η έννοια root (ρίζα) τοποθετείται στο κέντρο της απεικόνισης, οι υπό – έννοιες της ρίζας τοποθετούνται σε κύκλο γύρω από την έννοια ρίζα, κ.ο.κ.

Ο χρήστης μπορεί να επιλέξει μια έννοια κάνοντας click σε αυτή από την οπτική απεικόνιση.

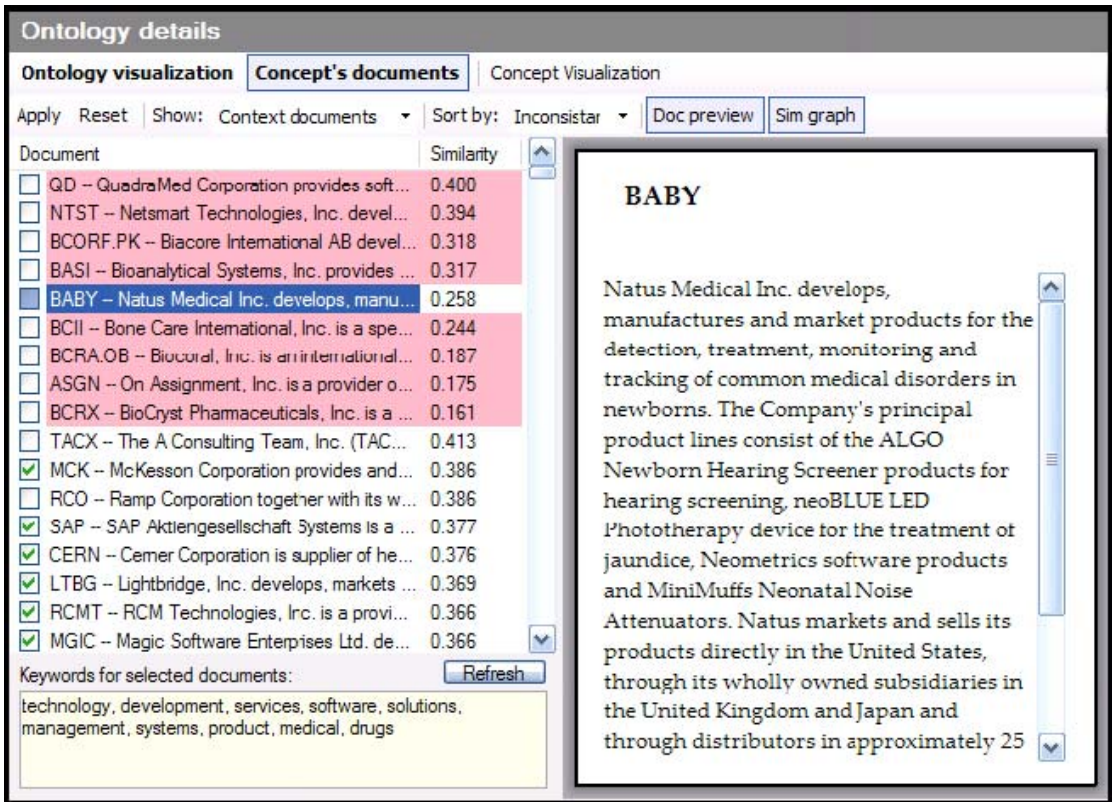
Η επιλεγμένη έννοια στη συγκεκριμένη χρονική στιγμή χρωματίζεται με κόκκινο χρώμα, οι άλλες έννοιες είναι με μπλε χρώμα και οι σχέσεις είναι με πράσινο χρώμα.

4.3.2 Λεπτομέρειες Έννοιας και Διαχείριση

Πληροφορίες και ιδιότητες μιας επιλεγμένης έννοιας παρουσιάζονται στο κάτω αριστερό μέρος του κύριου παράθυρου («Selected Concept’s details”,στο Σχήμα 4.1 και στην αριστερή πλευρά του Σχήματος 4.3).
Σε αυτό το μέρος, ο χρήστης μπορεί να αλλάξει το όνομα της επιλεγμένης έννοιας, να επιλέξει τις λέξεις κλειδιά της έννοιας και να δει τον αριθμό των στιγμιότυπων της έννοιας (All documents), τον αριθμό των στιγμιότυπων όχι σε οποιαδήποτε από τις υπό – έννοιες (Unused documents) και τη μέγιστη ομοιότητα των εγγράφων για τη συγκεκριμένη έννοια (Avg.Simil).



Σχήμα 4.3α: Μας δίνει λεπτομέρειες για την επιλεγμένη έννοια στον τρέχων χρόνο.



Σχήμα 4.3β: Η διεπαφή όπου γίνεται ο διαχείρισης των εννοιών.

Στο σύστημα υλοποιούνται δύο μέθοδοι εξαγωγής λέξεων κλειδιών. Η πρώτη μέθοδος, που παρουσιάζεται κάτω από τη λέξη *keywords,* συνθέτεται από τις λέξεις που είναι πιο περιγραφικές όσον αφορά το περιεχόμενο των στιγμιότυπων της έννοιας. Η δεύτερη μέθοδος, που παρουσιάζεται κάτω από τη φράση *SVM Keywords*,

αποτελείται από τις λέξεις οι οποίες είναι οι πιο διακριτικές για την επιλεγμένη έννοια σε σχέση με τις sibling έννοιες στην ιεραρχία. Εφόσον όμως η *SVM Keywords* μέθοδος είναι ακριβής σε χρόνο στο να εξάγει τα κλειδιά, χρησιμοποιείται μόνο όταν ο χρήστης το απαιτεί (με το πάτημα του κουμπιού «Calc»).

Μετά που μια νέα έννοια προστίθεται στην οντολογία, είτε με μη – επιβλεπόμενη είτε με επιβλεπόμενη μέθοδο, το σύστημα αυτόματα αναθέτει στιγμιότυπα σε αυτήν. Από τη στιγμή όμως που αυτή η ανάθεση συχνά δεν είναι τέλεια, ο χρήστης μπορεί με το χέρι (manually) να επαναθέσει στιγμιότυπα από και προς την έννοια διαμέσου της διεπαφής που φαίνεται στο Σχήμα 4.3β.

Η διεπαφή αποτελείται από τη λίστα των στιγμιότυπων, όπου τα στιγμιότυπα από την έννοια επιλέγονται. Η ανάθεση των στιγμιότυπων μπορεί εύκολα να αλλάξει, επιλέγοντας το τετραγωνάκι ή μη-επιλέγοντας το.

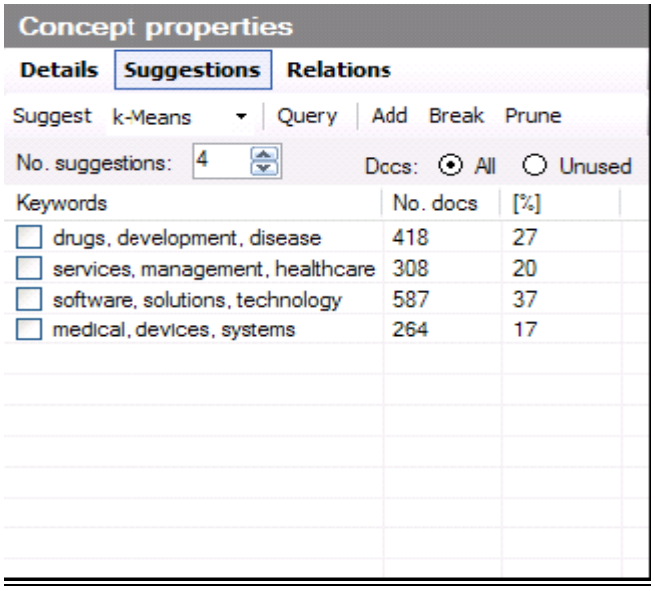
Το περιεχόμενο του επιλεγμένου στιγμιότυπου εμφανίζεται στη δεξιά πλευρά. Επίσης, το OntoGen έχει τη λειτουργικότητα να εντοπίζει συγκεκριμένες ασυνέπειες εντός της οντολογίας. Όταν ένα στιγμιότυπο δεν ανήκει στην επιλεγμένη έννοια αλλά είναι σε ένα από τα υπό – στιγμιότυπα, τότε χρωματίζεται έντονα το πλαίσιο του με κόκκινο χρώμα(για παράδειγμα όπως φαίνεται στο Σχήμα 4.3β).

4.4 Εισηγήσεις Εννοιών

Ένα από τα κύρια μέρη του συστήματος είναι οι εισηγήσεις των νέων πιθανών εννοιών που πρόκειται να προστεθούν στην οντολογία. Υπάρχουν δύο διαφορετικές προσεγγίσεις που υλοποιούνται για μάθηση εννοιών. Στη μη – επιβλεπόμενη προσέγγιση, το σύστημα παρέχει εισηγήσεις για πιθανές υπό – έννοιες για την επιλεγμένη έννοια. Στην επιβλεπόμενη προσέγγιση, ο χρήστης έχει μια αρχική ιδέα για το τι πρέπει να είναι μια υπό – έννοια (με το τι πρέπει να σχετίζεται), και την εισάγει στο σύστημα με μορφή ερώτησης (query).

4.4.1 Μη - Επιβλεπόμενη

Οι μη επιβλεπόμενες εισηγήσεις βασίζονται σε μεθόδους συσταδοποίησης από εξόρυξη κειμένων και τις ομάδες στιγμιότυπων από την επιλεγμένη έννοια τις αντιμετωπίζουμε σαν εισηγήσεις για τις υπό-έννοιες. Ο χρήστης μπορεί να επιβλέψει τον αριθμό των συστάδων (clusters), οι μέθοδοι συσταδοποίησης πρέπει να επιστρέφουν για ποια στιγμιότυπα πρέπει να εκτελείται η συσταδοποίηση (All για όλα τα στιγμιότυπα στην έννοια και Unused για τα στιγμιότυπα που δεν ανήκουν για την ώρα σε καμία από τις υπό – έννοιες της έννοιας). Ένα παράδειγμα εισηγήσεων φαίνεται στο Σχήμα 4.4. Ο χρήστης μπορεί να αποδεχτεί και να προσθέσει στην οντολογία οποιαδήποτε από τις παρεχόμενες εισηγήσεις, απλώς επιλέγοντας την και πατώντας το κουμπί «Add».

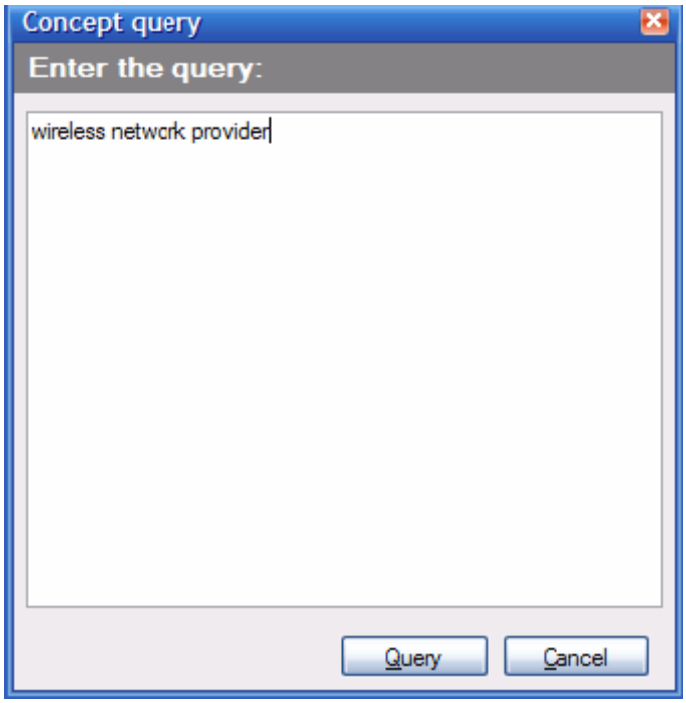


Σχήμα 4.4: Ο αλγόριθμος συσταδοποίησης ετοιμάζει εισηγήσεις και το πρόγραμμα τις τοποθετεί σε λίστα (όπως φαίνεται στο κάτω μισό της εικόνας). Κάθε εισηγημένη – προτεινόμενη υπό – έννοια περιγράφεται από τις εξαγόμενες κύριες λέξεις κλειδιά (χρησιμοποιώντας τη centroid μέθοδο) και το μέγεθος. Μια μεγαλύτερη λίστα από λέξεις κλειδιά εμφανίζεται, όταν ο χρήστης μετακινεί το ποντίκι πάνω στα Suggestions.

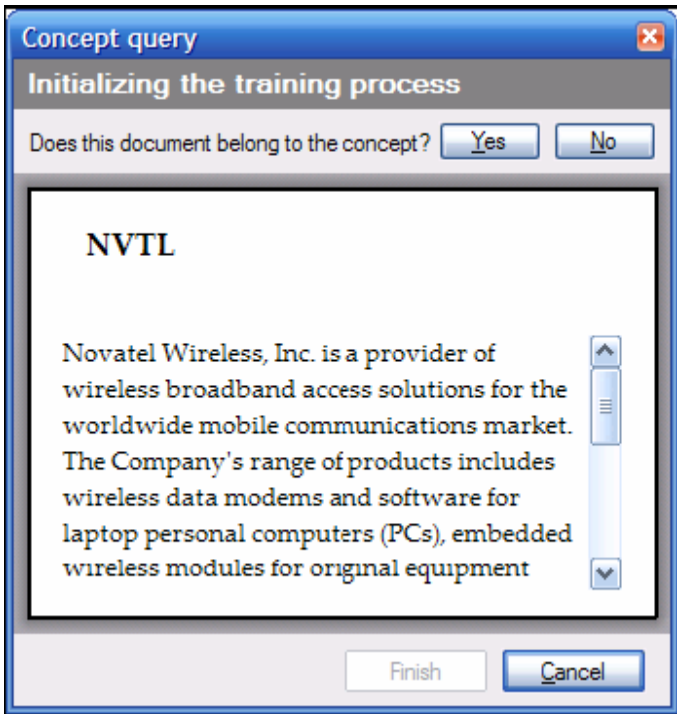
4.4.2 Επιβλεπόμενη

Ένα νέο χαρακτηριστικό στο OntoGen είναι μια επιβλεπόμενη μέθοδος για πρόσθεση εννοιών. Βασίζεται στην ενεργή μέθοδο μάθησης SVM και εφαρμόζεται μόνο στα στιγμιότυπα της που αφορούν την επιλεγμένη έννοια. Η επίβλεψη από τους χρήστες παρέχεται πρώτα μέσω ενός query (μορφή ερώτησης) ή ενός συνόλου από λέξεις κλειδιά που περιγράφουν την έννοια την οποία ο χρήστης έχει στο μυαλό και ακολουθείται από μια ακολουθία από ερωτήσεις (δέστε Σχήματα 4.5 α, β).

Σε κάθε βήμα το σύστημα ρωτά εάν ένα συγκεκριμένο στιγμιότυπο ανήκει στην έννοια και ο χρήστης μπορεί να επιλέξει Yes ή No. Οι ερωτήσεις επιλέγονται από τα στιγμιότυπα στα σύνορα μεταξύ αυτών που είναι πιο σχετικές με το query ή όχι και κατά συνέπεια αυτές που παρέχουν τις περισσότερες πληροφορίες στο σύστημα. Το σύστημα βελτιώνει μετά από κάθε απάντηση του χρήστη την προτεινόμενη έννοια, και ο χρήστης μπορεί να αποφασίσει πότε να σταματήσει τη διαδικασία αναλόγως με το πόσο ικανοποιημένος είναι με τις εισηγήσεις. Μετά που η έννοια κατασκευάζεται, προστίθεται στην οντολογία ως υπό – έννοια της επιλεγμένης έννοιας.



Σχήμα 4.5α: Ο χρήστης εισάγει μια πρόταση που να περιγράφει την νέα έννοια που επιθυμεί να προσθέσει.



Σχήμα 4.5β: Παράδειγμα ερώτησης από το σύστημα για το δεδομένο ερώτημα που εισήγαγε ο χρήστης προηγουμένως. Εμφανίζεται το στιγμιότυπο και η περιγραφή του, ώστε να βοηθήσει το χρήστη στην απόφαση του.

4.4.3 Παράδειγμα Μη-Επιβλεπόμενης και Επιβλεπόμενης Εισήγησης Έννοιας

Υπάρχει μια βασική διαφορά μεταξύ της μη επιβλεπόμενης και επιβλεπόμενης μεθόδου. Το κύριο πλεονέκτημα της μη επιβλεπόμενης μεθόδου, είναι ότι απαιτεί πολύ λίγα δεδομένα εισόδου από το χρήστη (αριθμός συστάδων). Η μη επιβλεπόμενη, παρέχει καλά ισορροπημένες εισηγήσεις για τις υπό – έννοιες που βασίζονται στα στιγμιότυπα και μπορεί να είναι πολύ χρήσιμες για εξερεύνηση δεδομένων.

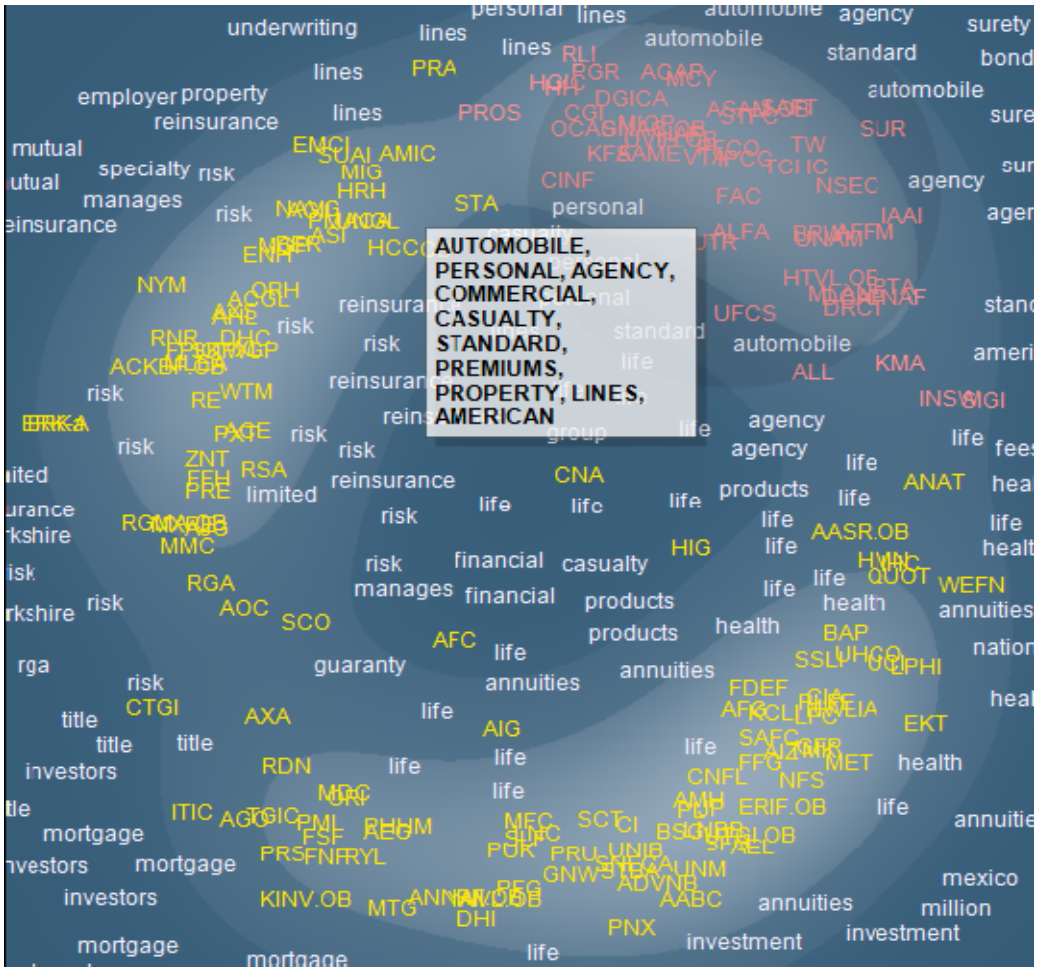
Από την άλλη πλευρά, η επιβλεπόμενη μέθοδος απαιτεί περισσότερα δεδομένα εισόδου. Ο χρήστης πρέπει αρχικά να υπολογίσει, να διαμορφώσει στη σκέψη του, τι πρέπει να είναι η υπό – έννοια, έπειτα πρέπει να περιγράψει την υπό – έννοια μέσω ενός ερωτήματος και να πάει διαμέσου της ακολουθίας των ερωτημάτων ούτως ώστε να διευκρινίσει το ερώτημα. Αυτό επιδιώκεται στις περιπτώσεις όπου ο χρήστης έχει μια ξεκάθαρη εικόνα σχετικά με τις υπό – έννοιες που θέλει να προσθέσει στην

οντολογία αλλά με τις μη επιβλεπόμενες μεθόδους αυτό δεν το ανακαλύπτει. Μια ιδέα για να τεθεί ένα ερώτημα μπορεί να προκύψει από την οπτική απεικόνιση της έννοιας που παρουσιάζεται παρακάτω.

4.5 Οπτική Απεικόνιση Έννοιας

Τα στιγμιότυπα μιας επιλεγμένης έννοιας μπορούν να απεικονιστούν οπτικά, χρησιμοποιώντας τεχνικές οπτικής απεικόνισης κειμένου – εγγράφου. Τα στιγμιότυπα παρουσιάζονται σε ένα χάρτη σαν σημεία, με τέτοιο τρόπο ώστε κάθε στιγμιότυπο που είναι παρόμοιο με κάποιο άλλο τοποθετείται κοντά σε αυτό, ενώ αντίθετα τοποθετείται μακριά το στιγμιότυπο που είναι λιγότερο παρόμοιο με το στιγμιότυπο που συγκρίνουμε. Η ομοιότητα μεταξύ των στιγμιότυπων υπολογίζεται βάση της κειμενικής περιγραφής τους, χρησιμοποιώντας την ομοιότητα συνημίτονου (καλά γνωστό μέτρο ομοιότητας στην εξόρυξη κειμένου και στην ανάκτηση πληροφοριών). Η πυκνότητα των στιγμιότυπων σε συγκεκριμένο μέρος του χάρτη, παρουσιάζεται από τη σύσταση του υποβάθρου. Οι περισσότερες κοινές λέξεις κλειδιά παρουσιάζονται για κάθε μια περιοχή του χάρτη και όταν ο χρήστης κινεί το ποντίκι γύρω από το χάρτη, μια εκτεταμένη λίστα από τις πιο κοινές λέξεις κλειδιά παρουσιάζεται για την περιοχή γύρω από το ποντίκι. Ο χρήστης μπορεί επίσης εστιάζοντας προς τα μέσα (zoom-in) να δει τις συγκεκριμένες περιοχές λεπτομερώς.

Η οπτική απεικόνιση είναι στενά ενσωματωμένη στο σύστημα. Ο χρήστης μπορεί να προσθέσει νέες υπό – έννοιες στην οντολογία, απλώς επιλέγοντας μια ομάδα στιγμιότυπων από το χάρτη. Αυτό είναι ιδιαίτερα χρήσιμο δεδομένου ότι η οπτική αναπαράσταση αντιπροσωπεύει συχνά τους σχηματισμούς ομάδων παρόμοιων στιγμιότυπων σε συστάδες. Υπάρχει ένα παράδειγμα στο Σχήμα 4.6.



Σχήμα 4.6: Η οπτική απεικόνιση δείχνει ένα χάρτη, όπου κάθε στιγμιότυπο είναι μια περιγραφή μιας ασφαλιστικής εταιρείας. Μπορούμε να δούμε ότι τα στιγμιότυπα ομαδοποιούνται ωραία σε τρείς ευδιάκριτες συστάδες: reassurance, home & car insurance, life insurance. Ο χρήστης επέλεξε τις περιπτώσεις από home & car insurance, χρησιμοποιώντας το ποντίκι (τα στιγμιότυπα σε ροζ χρώμα), έτσι ώστε να τις προσθέσει ως υπό – έννοια στην οντολογία.

4.6 Collaborative Editing and User Profiling

Το σύστημα προσφέρει επίσης τη συνεργάσιμη έκδοση των οντολογιών, όπου ο χρήστης μπορεί να χρησιμοποιήσει θέματα και σχέσεις που κατασκευάστηκαν προηγουμένως από άλλους χρήστες. Το σύστημα υποστηρίζει το χρήστη με την υποβολή παρόμοιων εισηγήσεων θεμάτων/σχέσεων από τη συλλογή των οντολογιών.

Η σκιαγράφηση χρηστών (user profiling), χρησιμοποιείται επίσης για να συντονίσει την αλληλεπίδραση ανθρώπου – οντολογίας, η οποία είναι βασισμένη στην προηγούμενη εργασία των χρηστών. Αυτό γίνεται με την καταγραφή των προηγούμενων επιλογών που έκανε ο χρήστης κατά τη διάρκεια κατασκευής οντολογιών και επιπλέον για τη χρησιμοποίηση τους ως ένα επιπρόσθετο δεδομένο που παρέχει μια εξατομικευμένη άποψη σχετικά με τα υποκείμενα δεδομένα και την οντολογία που έχει κατασκευαστεί μέχρι στιγμής, μέσω "personalized word weighting schemas" (αντί της χρησιμοποίησης των προκαθορισμένων σχημάτων, όπως το TFIDF).

4.7 Συμπεράσματα

Στο πιο πάνω κεφάλαιο έχω παραθέσει τον «ημι-αυτόματο» όπως συχνά χαρακτηρίζεται από τους κατασκευαστές του, συντάκτη οντολογιών όπως εμφανίζεται στην νέα έκδοση του συστήματος OntoGen. Θα το δούμε αναλυτικότερα στο κεφάλαιο της υλοποίησης. Αυτό που θέλω να θίξω κυρίως κλείνοντας αυτό το κεφάλαιο, είναι ότι παρόλο που το εργαλείο αυτό βασίζεται σε εκτενή χρήση της μηχανικής μάθησης και των μεθόδων εξόρυξης κειμένων, τα οποία βοηθούν το χρήστη στα βασικά βήματα της διαδικασίας κατασκευής οντολογιών, εντούτοις το OntoGen μπορεί να χρησιμοποιηθεί από άτομα τα οποία δεν διαθέτουν κανένα υπόβαθρο στους τομείς οι οποίοι αναφέρθηκαν πιο πάνω. Και αυτό είναι κάτι το οποίο αποδείχτηκε κατά την αξιολόγηση του συστήματος, κάτι που θα δούμε επίσης σε μετέπειτα κεφάλαιο.

Κεφάλαιο 5

ΑΛΓΟΡΙΘΜΟΙ ΣΤΟΥΣ ΟΠΟΙΟΥΣ ΣΤΗΡΙΖΕΙ

ΤΗ ΛΕΙΤΟΥΡΓΙΑ ΤΟΥ ΤΟ ΣΥΣΤΗΜΑ ONTOGEN

5.1 Εισαγωγή

Είδαμε στο προηγούμενο κεφάλαιο της περιγραφής και ανάλυσης του συστήματος OntoGen, το σύστημα να παρουσιάζεται ως ένας ημι – αυτόματος κατασκευαστής θεματικών οντολογιών. Το σύστημα αλληλεπιδρά με το χρήστη και τον βοηθά κατευθύνοντας τον σε διάφορα σημεία, κατά τη διάρκεια της κατασκευής της οντολογίας. Εισηγείται έννοιες, σχέσεις και ονομασίες για αυτές, αναθέτει αυτόματα

στιγμιότυπα στις έννοιες και παρέχει στο χρήστη μια καλή επισκόπηση της οντολογίας διαμέσου της οπτικής απεικόνισης και concept browsing.

Σε αυτό το κεφάλαιο πρόκειται να αναφερθούν οι κυριότεροι αλγόριθμοι που υλοποιούνται στο σύστημα που περιγράφηκε προηγουμένως και τι λειτουργία εξυπηρετεί ο καθένας από αυτούς.

5.2 Τεχνικές που ακολουθούνται προς ανακάλυψη θέματος σε συλλογές κειμενικών εγγράφων

Όταν ασχολούμαστε με μεγάλες συλλογές εγγράφων είναι δύσκολο να κατανοήσουμε και να επεξεργαστούμε όλες τις πληροφορίες που περιέχονται σε αυτές. Τυποποιημένες τεχνικές εξόρυξης κειμένου και ανάκτησης πληροφοριών συνήθως στηρίζονται στο ταίριασμα λέξης (word matching) και δεν λαμβάνουν υπόψη την ομοιότητα των εγγράφων και τη δομή τους μέσα στη συλλογή. Προσπαθούμε να το υπερνικήσουμε αυτό, με το να εξάγουμε αυτόματα τα θέματα που καλύπτουν τα έγγραφα μέσα στις συλλογές και βοηθώντας το χρήστη να τα οργανώσει σε μια θεματική οντολογία.

Η θεματική οντολογία είναι ένα σύνολο από θέματα που ενώνονται με διαφορετικούς τύπους από σχέσεις. Κάθε θέμα συμπεριλαμβάνει ένα σύνολο από συσχετιζόμενα έγγραφα.

Η δημιουργία μιας τέτοιας οντολογίας από μια δεδομένη συλλογή, μπορεί να είναι πολύ χρονοβόρα για το χρήστη. Προκειμένου να πάρει μια πρώτη αίσθηση ο χρήστης τι είδους θέματα βρίσκονται στη συλλογή, ποιες είναι οι σχέσεις μεταξύ των θεμάτων και να αναθέσει κάθε έγγραφο με κάποιο συγκεκριμένο θέμα, ο χρήστης πρέπει να πάει διαμέσου όλων των εγγράφων και να τα επεξεργαστεί όλα με το χέρι. Αυτό πάλι προσπάθησε να υπερνικηθεί με τη δημιουργία του OntoGen, το οποίο όπως ξαναείπαμε είναι ένα ειδικό εργαλείο που βοηθά το χρήστη προτείνοντας του πιθανά νέα θέματα και εμφανίζοντας του οπτικά τη θεματική οντολογία σε πραγματικό χρόνο.

Στο σύστημα μας έχουμε δύο διαφορετικές προσεγγίσεις για ανακάλυψη θεμάτων μέσα από μια συλλογή εγγράφων. Δύο διαφορετικούς αλγόριθμους που θα περιγραφούν πιο κάτω, οι οποίοι χρησιμοποιούνται για την αυτόματη εισήγηση θεμάτων κατά τη διάρκεια της κατασκευής της θεματικής οντολογίας.

5.3 Latent Semantic Indexing (LSI)

Αυτή η πρώτη προσέγγιση είναι μια γραμμική τεχνική μείωσης των διαστάσεων. Η τεχνική στηρίζεται στο γεγονός ότι οι λέξεις που σχετίζονται με το ίδιο θέμα εμφανίζονται μαζί συχνότερα από τις λέξεις που σχετίζονται σε διαφορετικά θέματα. Το αποτέλεσμα της τεχνικής LSI είναι συγκεχυμένες συστάδες λέξεων, όπου η κάθε μια περιγράφει ένα θέμα.

Η γλώσσα περιέχει πολλές περιττές πληροφορίες, μιας και πολλές λέξεις έχουν κοινό ή παρόμοιο νόημα. Αυτό μπορεί να είναι δύσκολο για τον υπολογιστή για να το χειριστεί χωρίς κάποια επιπλέον πληροφορία – γνώση υποβάθρου. Η LSI, είναι μια τεχνική για εξαγωγή γνώσης υποβάθρου από τα κειμενικά έγγραφα. Χρησιμοποιεί μια τεχνική από τη γραμμική άλγεβρα που λέγεται Singular Value Decomposition (SVD) και την αναπαράσταση bag-of-words για τα κειμενικά έγγραφα για την ανίχνευση λέξεων με παρόμοιο νόημα. Αυτό μπορεί να αντιμετωπισθεί επίσης και σαν εξαγωγή των κρυμμένων σημασιολογικών εννοιών ή θεμάτων από τα κειμενικά έγγραφα.

Η LSI υπολογίζεται ως ακολούθως. Πρώτα ο όρος-έγγραφο πίνακας A κατασκευάζεται από ένα δοθέν σύνολο από κειμενικά έγγραφα. Αυτός είναι ένας πίνακας με bag-of-words διανύσματα εγγράφων που εμφανίζονται σαν στήλες. Αυτός ο πίνακας αναλύεται χρησιμοποιώντας SVD έτσι ώστε $A=USV^T$, όπου οι πίνακες U και V είναι ορθογώνιοι και ο S είναι διαγώνιος πίνακας με διατεταγμένες μοναδικές τιμές στη διαγώνιο. Οι στήλες στον πίνακα U σχηματίζουν μια ορθογώνια βάση του υπό-διαστήματος σε ένα bag-of-words διάστημα και τα διανύσματα με τα τις ψηλότερες μοναδικές τιμές φέρουν και τις περισσότερες πληροφορίες. Με βάση αυτό μπορούμε να δούμε τα διανύσματα τα οποία αποτελούν τη βάση σαν έννοιες ή

θέματα. Το διάστημα που εκτείνεται από αυτά τα διαστήματα λέγεται Semantic Space.

5.4 K-Means clustering

Η δεύτερη προσέγγιση που χρησιμοποιείται από το σύστημα για εξαγωγή θεμάτων είναι ο αλγόριθμος συσταδοποίησης k-means. Ο αλγόριθμος χωρίζει τη συλλογή εγγράφων που έχουμε κάθε φορά σε k clusters (συστάδες, ομάδες), έτσι ώστε δύο έγγραφα που περιέχονται μέσα στην ίδια συστάδα συσχετίζονται περισσότερο από ότι δύο έγγραφα από δύο διαφορετικές συστάδες.

Η συσταδοποίηση είναι μια τεχνική για διαχωρισμό των δεδομένων έτσι ώστε κάθε συστάδα (cluster) να περιέχει μόνο τα σημεία που είναι παρόμοια σύμφωνα με κάποιο προκαθορισμένο μετρικό. Στη περίπτωση που έχουμε κείμενο, αυτό υλοποιείται με την εύρεση ομάδων με κοινά έγγραφα, τα οποία είναι έγγραφα που μοιράζονται παρόμοιες λέξεις.

Ο k-means αλγόριθμος είναι ένας επαναληπτικός αλγόριθμος ο οποίος χωρίζει τα δεδομένα σε k συστάδες. Έχει ήδη χρησιμοποιηθεί επιτυχώς στα έγγραφα κειμένων για να ομαδοποιήσει μια μεγάλη συλλογή εγγράφων βασισμένη στο θέμα των εγγράφων και που ενσωματώνει μια προσέγγιση για την οπτική απεικόνιση αυτής της μεγάλης συλλογής εγγράφων. Πιο κάτω μπορείτε να δείτε τον αλγόριθμο κατά προσέγγιση.

Algorithm : K-Means.

Input: A set of data points, a distance metric, the desired number
of clusters k
Output: Clustering of the data points into k clusters
(1) Set k cluster centers by randomly picking k data points as
cluster centers
(2) **repeat**
(3) Assign each point to the nearest cluster center

(4) Recompute the new cluster centers
(5) **until** the assignment of data points has not changed

5.5 Μέθοδοι για Εξαγωγή των Κυρίως Λέξεων Κλειδιών

Το σύστημα παρέχει επίσης δύο μεθόδους για εξαγωγή των κύριων λέξεων κλειδιών, οι οποίες βοηθούν το χρήστη να κατανοήσει και να ονομάσει τα θέματα. Δηλαδή καλύτερα, για να παράξουν περιγραφή για ένα δεδομένο θέμα το οποίο βασίζεται στα έγγραφα του θέματος. Αυτές είναι η εξαγωγή κλειδιού χρησιμοποιώντας Centroid Vectors (περιγραφικές λέξεις κλειδιά) και η εξαγωγή κλειδιού χρησιμοποιώντας Support Vector Machine (διακριτικές λέξεις κλειδιά).

Τα κειμενικά έγγραφα αναπαριστούνται στο σύστημα σαν διανύσματα. Χρησιμοποιούμε μια τυποποιημένη προσέγγιση τη Bag-of-words(BOW) μαζί με την TFIDF weighting. Η ομοιότητα μεταξύ δύο εγγράφων ορίζεται ως το συνημίτονο της γωνίας μεταξύ των αναπαραστάσεων των διανυσμάτων τους – cosine similarity.

5.6 Εξαγωγή Κλειδιού χρησιμοποιώντας Centroid Vectors

Λέγοντας «centroid» εννοούμε το άθροισμα όλων των διανυσμάτων των εγγράφων μέσα στο θέμα. Οι κύριες λέξεις κλειδιά επιλέγονται να είναι οι λέξεις με τα μεγαλύτερα βάρη στο centroid vector.

5.7 Εξαγωγή Κλειδιού χρησιμοποιώντας Support Vector Machine (SVM)

Έστω Α το θέμα που θέλουμε να περιγράψουμε με λέξεις κλειδιά. Παίρνουμε όλα τα έγγραφα από τα θέματα τα οποία έχουν ως υπό – θέμα το Α και σημαδεύουμε αυτά τα έγγραφα ως αρνητικά. Παίρνουμε όλα τα έγγραφα το θέμα Α και τα σημαδεύουμε ως θετικά. Εάν ένα έγγραφο είναι σημαδεμένο και θετικό και αρνητικό, τότε λέμε ότι είναι θετικό. Κατόπιν μαθαίνουμε ένα γραμμικό ταξινομητή SVM πάνω σε αυτά τα έγγραφα και κατηγοριοποιεί τα centroid του θέματος Α. Οι λέξεις κλειδιά που

περιγράφουν την έννοια Α είναι λέξεις, των οποίων τα βάρη στο κανονικό διάνυσμα SVM συμβάλλουν το περισσότερο όταν θα αποφασιστεί εάν το centroid είναι θετικό (δηλ. ανήκει στο θέμα).

Η διαφορά μεταξύ των δύο προσεγγίσεων είναι ότι η δεύτερη προσέγγιση λαμβάνει υπόψη το πλαίσιο του θέματος. Ας υποθέσουμε ότι έχουμε ένα θέμα που λέγεται « υπολογιστές». Όταν αποφασίζουμε ποιες θα είναι οι λέξεις κλειδιά για κάποιο υπό – θέμα Α, η πρώτη μέθοδος θα κοιτάξει μόνο ποιες είναι οι ποιο σημαντικές λέξεις μέσα στο υπό – θέμα Α και λέξεις όπως «υπολογιστής» θα βρεθούν πιθανότατα ως οι πιο σημαντικές. Ωστόσο, γνωρίζουμε ήδη ότι το Α είναι ένα υπό – θέμα του « υπολογιστές» και ενδιαφερόμαστε περισσότερο να βρούμε τις λέξεις κλειδιά οι οποίες τις διαχωρίζουν από τα άλλα έγγραφα στο θέμα « υπολογιστές». Η δεύτερη μέθοδος το κάμνει αυτό, παίρνοντας τα έγγραφα από όλα τα υπέρ – θέματα του Α σαν πλαίσιο και μαθαίνει τις πιο κρίσιμες λέξεις χρησιμοποιώντας τον SVM.

5.8 Αναπαράσταση εγγράφων

Όπως αναφέρθηκε πιο πάνω τα έγγραφα αναπαριστώνται στο σύστημα με τη συνηθέστερα χρησιμοποιούμενη μέθοδο αναπαράστασης εγγράφων, την bag-of-words. Κάθε έγγραφο κωδικοποιείται ως ένα διάνυσμα από όρους συχνοτήτων και η ομοιότητα ενός ζεύγους εγγράφων υπολογίζεται από τον αριθμό και το βάρος των λέξεων οι οποίες μοιράζονται σε αυτά τα δύο έγγραφα.

Έστω L = {w1,…,wn} ένα λεξιλόγιο από λέξεις. Θέτουμε το TFk να είναι ο αριθμός που συναντούμε τη λέξη wk στο έγγραφο. Στην αναπαράσταση bag-of-words ένα έγγραφο μόνο του, κωδικοποιείται ως ένα διάνυσμα χ με στοιχεία που αντιστοιχούν στις λέξεις από ένα λεξιλόγιο παρέχοντας κάποιο βάρος στις λέξεις,

π.χ. $x^k = TF_k$.

Το μέτρο που χρησιμοποιείται συνήθως για να συγκρίνει συνήθως
τα έγγραφα κειμένων είναι το cosine similarity και ορίζεται να είναι το συνημίτονο της γωνίας μεταξύ δύο διανυσμάτων σε μορφή bag-of-words.

$$sim(x_i, x_j) = \sum_{k=1}^{n} x_i^k \cdot x_j^k \Big/ \sqrt{\sum_{k=1}^{n} x_i^k \cdot x_i^k} \sqrt{\sum_{k=1}^{n} x_j^k \cdot x_j^k} \cdot$$

Η απόδοση του bag-of-words καθώς και του cosine similarity μπορεί να επηρεαστεί σημαντικά από τα βάρη των λέξεων. Σε κάθε λέξη από το λεξιλόγιο V ανατίθεται ένα βάρος και τα διανύσματα xi πολλαπλασιάζονται με τα αντίστοιχα βάρη.

Κάτι ακόμα που δεν έχουμε αναφέρει σχετικά με τον υπολογισμό του βάρους των λέξεων, ο οποίος γίνεται από το λεγόμενο TFIDF weighting.

5.9 Συμπέρασμα

Συνοπτικά αναφέραμε τις διάφορες προσεγγίσεις και τεχνικές που ακολουθούνται στα διάφορα σημεία του συστήματος ούτως ώστε να έχουμε μια πιο ξεκάθαρη άποψη. Στο σύστημα λοιπόν ενσωματώνονται μέθοδοι μηχανικής μάθησης καθώς και τεχνικές εξόρυξης κειμένου. Ακόμη για την ανακάλυψη εννοιών ακολουθούνται δύο μέθοδοι, μη – επιβλεπόμενοι και επιβλεπόμενοι. Στις μη-επιβλεπόμενες μεθόδους εμπίπτουν το LSI και το k-means clustering. Στις επιβλεπόμενες έχουμε την ενεργή μάθηση και την οπτική απεικόνιση των εννοιών. Μέθοδοι που βοηθούν στην κατανόηση των ανακαλυφθέντων εννοιών, είναι η εξαγωγή λέξεων κλειδιών και η οπτική απεικόνιση των εννοιών. Στην εξαγωγή λέξεων κλειδιών έχουμε τη μέθοδο TFIDF και τη SVM-normal. Στην Απεικόνιση εννοιών χρησιμοποιείται η LSI και πολύ-διάστατο scaling και επίσης παρέχεται και ως ξεχωριστό εργαλείο, γνωστό ως Document Atlas (**http://docatlas.ijs.si**)

Κεφάλαιο 6

ΤΟ ONTOGEN ΣΤΗΝ ΠΡΑΞΗ

6.1 Εισαγωγή

Στα προηγούμενα κεφάλαια περιγράψαμε τα κύρια συστατικά του συστήματος OntoGen και τις τεχνικές εξόρυξης κειμένου που κρύβονται πίσω από αυτό. Εδώ θα δείξουμε πως αυτά τα συστατικά συνδυάζονται σε μια απλή διαδικασία της δημιουργίας μιας θεματικής οντολογίας. Στη συνέχεια του κεφαλαίου θα παρουσιαστεί το σχέδιο που ακολουθήθηκε και τα αποτελέσματα που εξήχθηκαν από μια μελέτη χρηστών προκειμένου να αξιολογηθεί η διαδικασία παραγωγής οντολογίας. Η μελέτη έχει εφαρμοστεί στο εργαλείο που μελετάμε φυσικά, κατασκευής οντολογιών, το OntoGen. Η σε βάθος ανάλυση της εμπειρίας του χρήστη, παραδίδει πολύτιμες ιδέες σε ότι αφορά τις απαιτήσεις για την περαιτέρω ανάπτυξη του συστήματος.

6.2 Παράδειγμα θεματικής οντολογίας

Σε αυτό το κεφάλαιο θα δείξουμε ένα παράδειγμα δημιουργίας θεματικής οντολογίας (topic ontology) από 7177 περιγραφές εταιρειών, που πάρθηκαν από το Yahoo!Finance (http://finance.yahoo.com/). Κάθε εταιρεία περιγράφεται από μια παράγραφο κειμένου. Μια τυπική περιγραφή παρμένη από το Yahoo! θα έχει τη μορφή που φαίνεται παρακάτω:

YAHOO! INC. IS A PROVIDER OF INTERNET PRODUCTS AND SERVICES TO CONSUMERS AND BUSINESSES THROUGH THE YAHOO! NETWORK, ITS WORLDWIDE NETWORK OF ONLINE PROPERTIES. THE COMPANY'S PROPERTIES AND SERVICES FOR CONSUMERS AND BUSINESSES RESIDE IN FOUR AREAS: SEARCH AND MARKETPLACE, INFORMATION AND CONTENT, COMMUNICATIONS AND CONSUMER SERVICES AND AFFILIATE SERVICES...

Χρησιμοποιώντας το OntoGen σε αυτές τις περιγραφές, κάποιος μέσα σε πολύ λίγο χρόνο (στη δική μας περίπτωση γύρω στα 15 λεπτά!), μπορεί να δημιουργήσει μια οντολογία από περιοχές με εκείνες τις εταιρείες τις οποίες καλύπτει το Yahoo!Finance.

Παρακάτω θα δούμε τα βασικά βήματα που ακολουθούνται για τη δημιουργία της συγκεκριμένης οντολογίας, όπου σε κάθε βήμα θα εμφανίζουμε παράλληλα και τις αντίστοιχες οθόνες του συστήματος μας έτσι ώστε να έχουμε μια καλύτερη αίσθηση της όλης διαδικασίας που ακολουθείται.

6.3 Βασικά Βήματα Υλοποίησης Μέσα Από Οθόνες

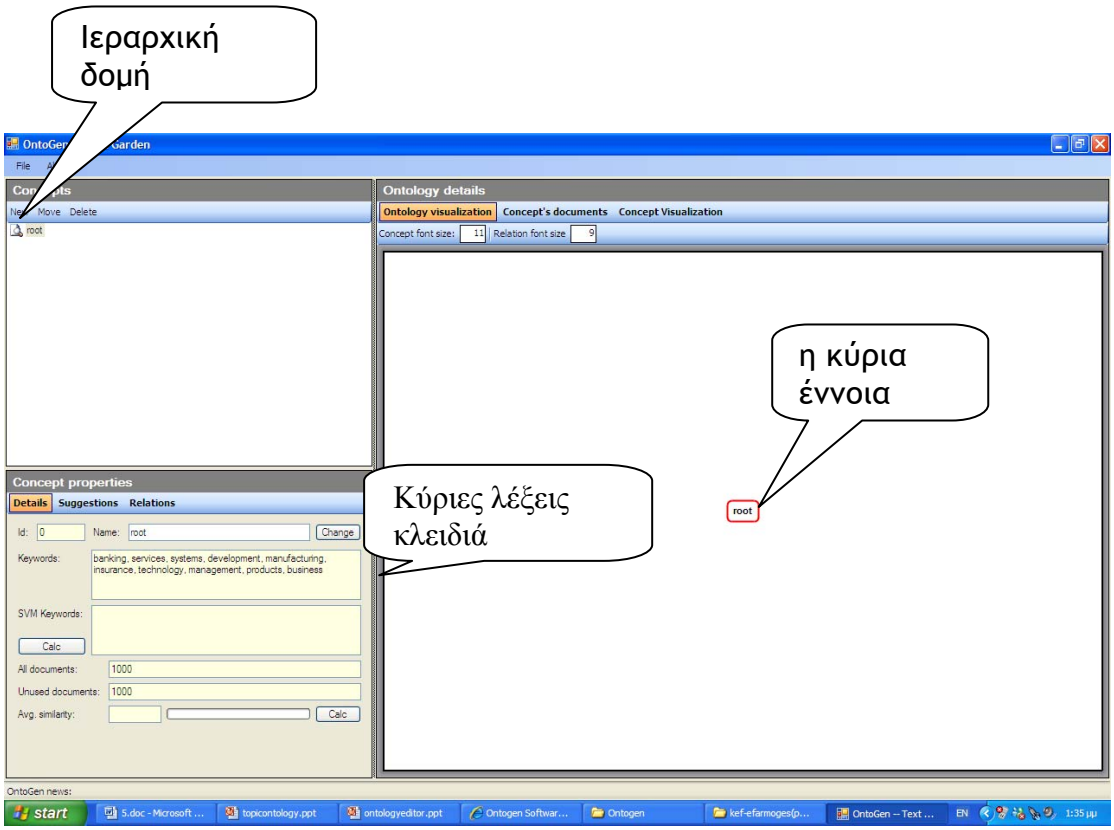
Το κύριο παράθυρο της διεπαφής όπως αναφέρθηκε και σε προηγούμενα κεφάλαια, χωρίζεται σε τρεις περιοχές με την μεγαλύτερη περιοχή του παραθύρου να χρησιμοποιείται για την οπτική αναπαράσταση της οντολογίας(δεξιά) και τη διαχείριση των εγγράφων(αριστερό μέρος). Στο πάνω αριστερό μέρος είναι το εννοιολογικό δένδρο, όπου εκεί μπορούμε να δούμε όλες τις έννοιες που υπάρχουν στην οντολογία μας και κάτω αριστερά, είναι το μέρος όπου ο χρήστης μπορεί να

ελέγξει τις λεπτομέρειες και να διαχειριστεί τα σχετικά που αφορούν μια επιλεγμένη έννοια από το σχήμα και να πάρει εισηγήσεις (suggestions) για τις υπό-έννοιες.

Αυτά αναφέρονται απλώς για να θυμηθούμε τους βασικούς χώρους στο σύστημα μας, τώρα όμως προχωρούμε με την υλοποίηση.

Αρχικά λοιπόν στο σύστημα εισάγουμε ή καλύτερα φορτώνουμε τη βάση δεδομένων από κείμενα πάνω στην οποία θα γίνει η υλοποίηση, δηλαδή το κτίσιμο της οντολογίας.
Να επισημάνουμε εδώ ότι το OntoGen υποστηρίζει διάφορες μορφές κειμένων για είσοδο (Folder and Line-Documents) και υποστηρίζει επίσης και Bag-Of-Words μορφή, που είναι αυτή η μορφή κειμένων που θα δούμε εδώ.

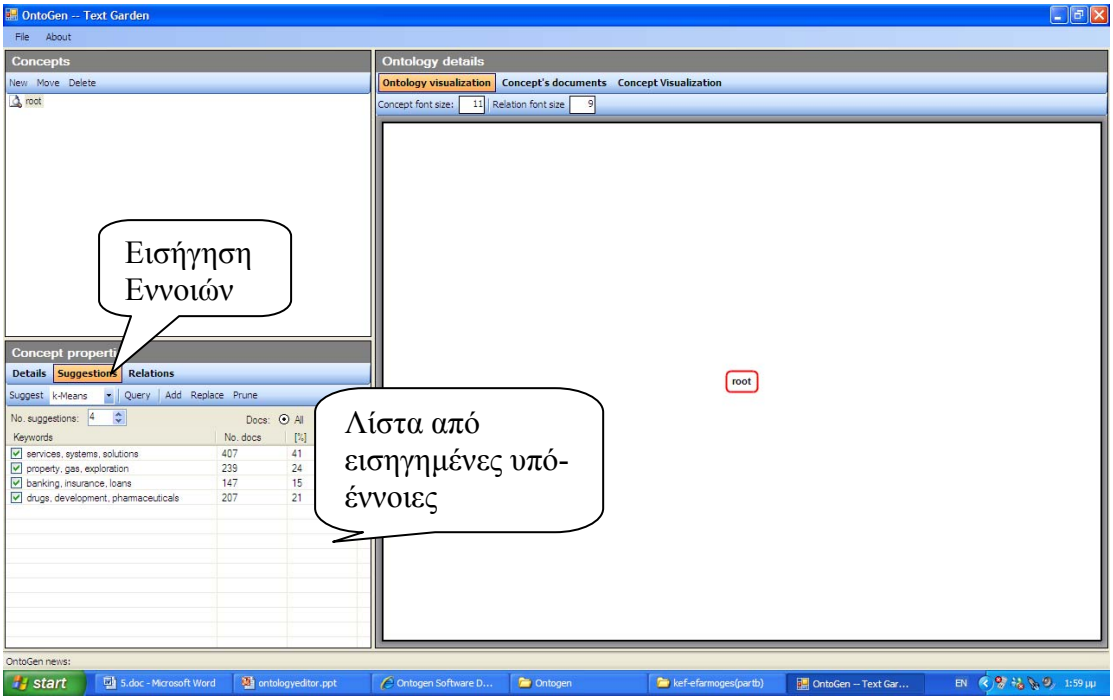
Έτσι μόλις η βάση φορτωθεί (τα κείμενα), έχουμε την παρακάτω αναπαράσταση. Δηλαδή τη δημιουργία της root concept τα οποία ο Blanz Fortuna ένας εκ των τριών δημιουργών του συστήματος αυτού, περιγράφει το root, “this is everything”.
Αρχικά σε αυτή την έννοια περιέχονται όλα τα κείμενα της βάσης δεδομένων που έχουμε φορτώσει στο σύστημα, αφού προς το παρόν κανένας αλγόριθμος δεν έχει εφαρμοστεί, καμία επεξεργασία δεν έχει γίνει ακόμα στην είσοδο μας. (Σχήμα 6.1)



Σχήμα 6.1: Δημιουργία της κύριας έννοιας (root concept) .

Με το φόρτωμα της βάσης, δημιουργείται η έννοια root, εμφανίζονται επίσης οι κύριες λέξεις κλειδιά αυτής της βάσης κειμένων που έχουμε φορτώσει στο σύστημα, όπως φαίνεται στο αριστερό κάτω μέρος και πάνω αριστερά το σύστημα τοποθετεί ιεραρχικά τις έννοιες που έχουν δημιουργηθεί μέχρι αυτή την ώρα(το root μέχρι στιγμής).

Στη συνέχεια, μπορούμε να αρχίσουμε να προσθέτουμε έννοιες στο σύστημα.



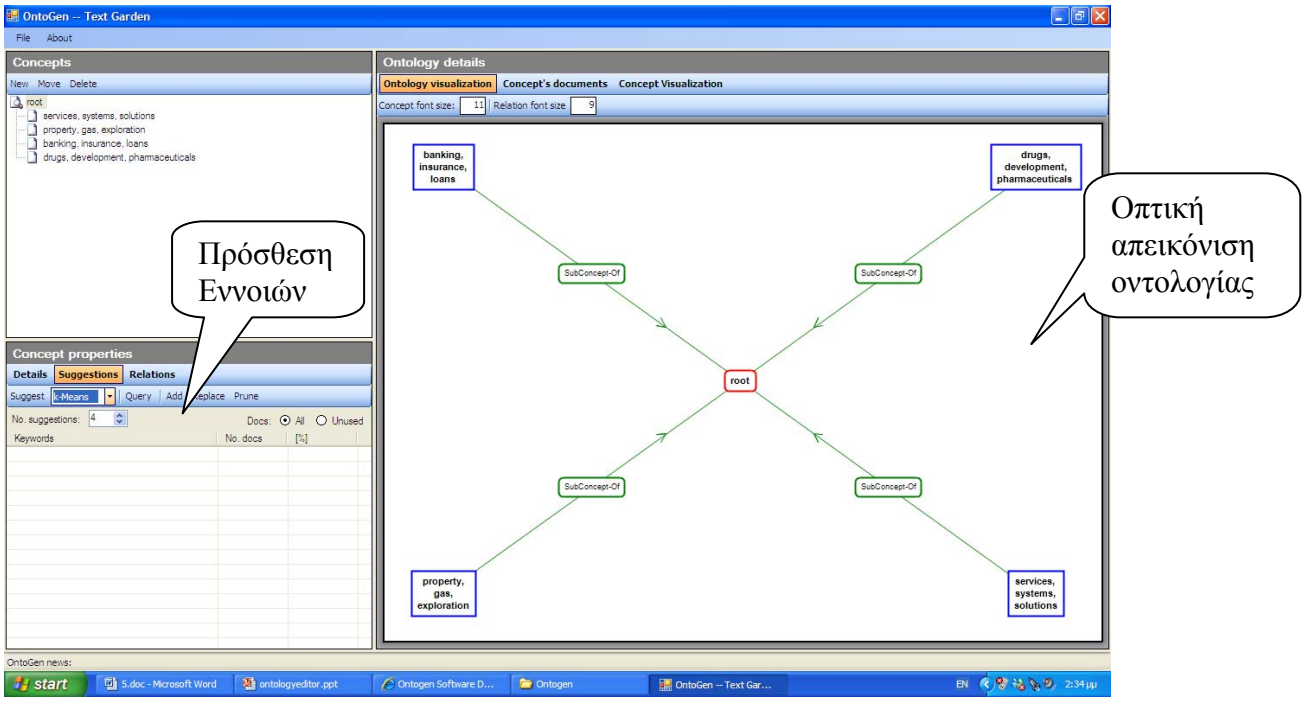
Σχήμα 6.2: Εισήγηση Εννοιών από το σύστημα με τη μορφή λίστας .

Αυτό το κάμνει, πατώντας στο Suggestions όπως φαίνεται στην Σχήμα 6.2. Τότε στο μέρος αυτό του συστήματος μπορούμε να προσθέσουμε πιθανές υπό-έννοιες στην οντολογία μας μέσω κάποιων εισηγήσεων (έτσι τις ονομάζει) που το σύστημα μας κάμνει για κάποια έννοια την οποία επιλέξαμε.

Στο σύστημα μπορούμε να επιλέξουμε τη μέθοδο με την οποία θέλουμε να γίνει η κατηγοριοποίηση στα δεδομένα μας. Έστω επιλέγουμε τον k-means αλγόριθμο και αριθμό εισηγήσεων = 4, έτσι προκύπτει μια λίστα από τέσσερις πιθανές έννοιες (Σχήμα 6.2).

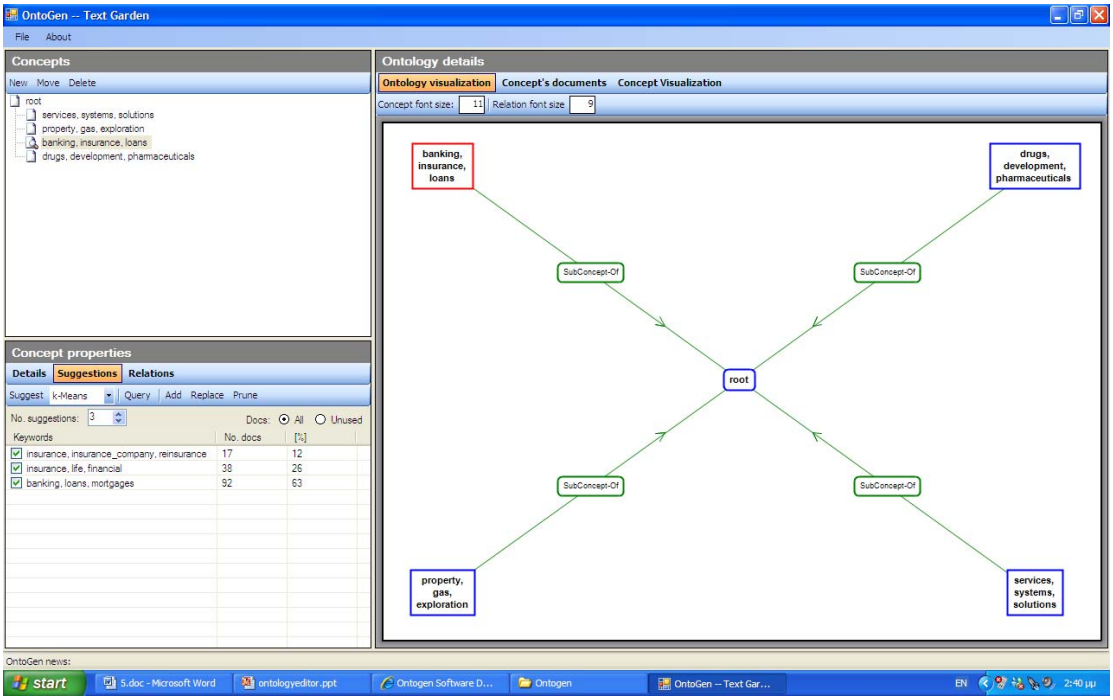
Ο χρήστης μπορεί να παίξει με τις παραμέτρους, να αλλάξει για παράδειγμα τον αριθμό των εισηγήσεων ή να μην επιλέξει κάποια από τις ομάδες που δίνει η κατηγοριοποίηση. Γι αυτό άλλωστε και το σύστημα χαρακτηρίστηκε semi-automatic, επειδή πάντοτε η τελική απόφαση αφήνεται στις επιλογές του χρήστη!

Αφού λοιπόν ο χρήστης επιλέξει από τη λίστα που προκύπτει ποιες θα είναι οι νέες έννοιες που θα προστεθούν στην οντολογία, πατάει «Add» και έχουμε το παρακάτω Σχήμα. (Σχήμα 6.3)



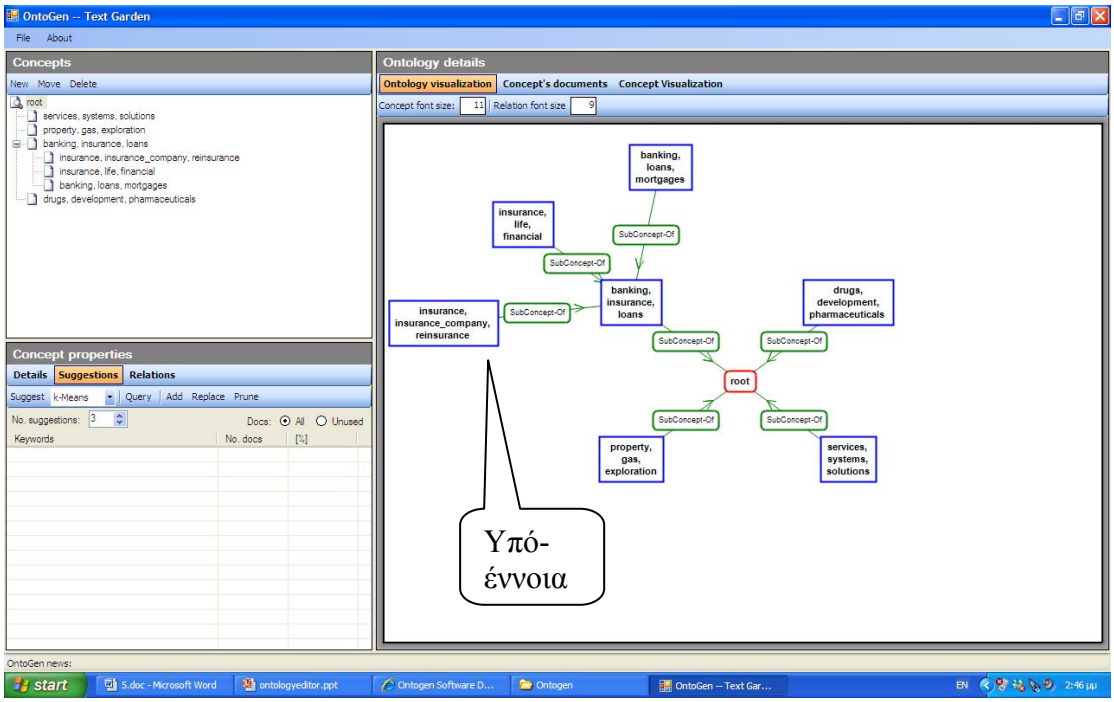
Σχήμα 6.3: Οι έννοιες που επέλεξε ο χρήστης από τη λίστα των προτεινόμενων Εισηγήσεων προστίθενται στην οντολογία και τις βλέπουμε και οπτικά όπως φαίνονται στο δεξιό παράθυρο.

Μετά επιλέγουμε μια έννοια από την Οπτική απεικόνιση της οντολογίας, έστω την έννοια: banking, insurance, loans.



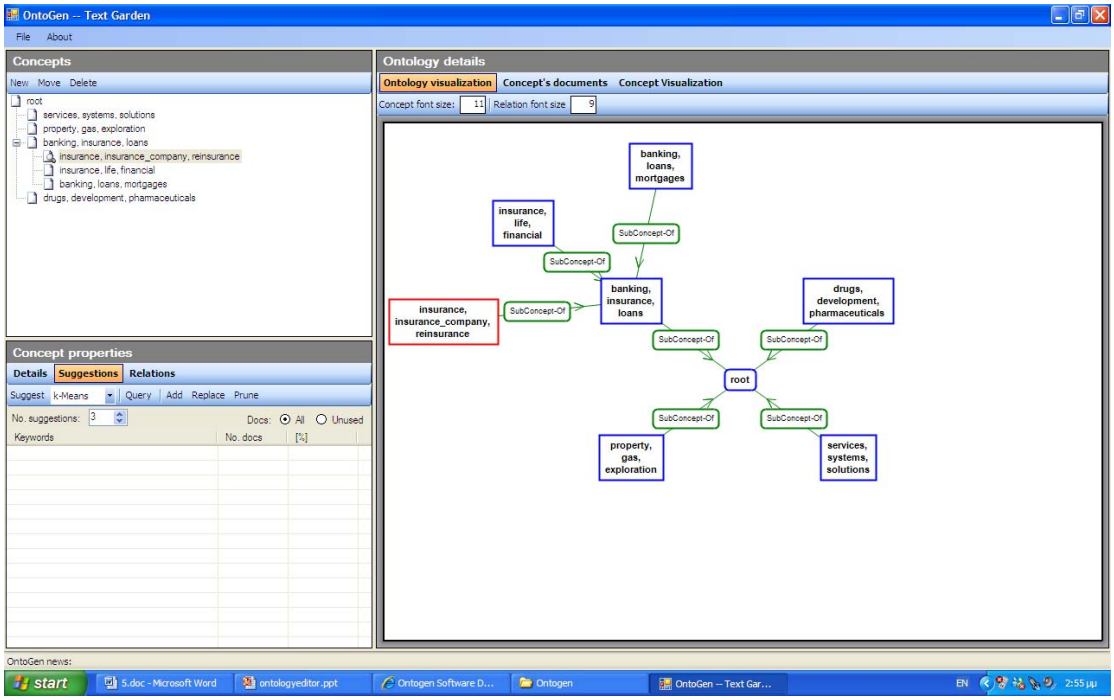
Σχήμα 6.4: Επιλέγουμε μια έννοια από την οπτική αναπαράσταση της οντολογίας για να προσθέσουμε νέες υπό- έννοιες σε αυτήν. Η επιλεγμένη έννοια φαίνεται με κόκκινο χρώμα.

Η επιλεγμένη έννοια εμφανίζεται με κόκκινο χρώμα όπως φαίνεται και στο Σχήμα 6.4. Έπειτα ακολουθούμε την ίδια διαδικασία όπως προηγουμένως, δηλαδή με τον αλγόριθμο k-means και επιλέγοντας αυτή τη φορά 3 για αριθμό εισηγήσεων, πάμε να προσθέσουμε υπό – έννοιες στην έννοια αυτή που επιλέξαμε. Μας εμφανίζει την λίστα και πάλι έχοντας ο χρήστης την τελική κρίση αποφασίζει ποιες υπό- έννοιες θα προσθέσει στην οντολογία από αυτές που το σύστημα του εξάγει αυτόματα. Και με το πάτημα του κουμπιού «Add» έχουμε αυτόματα το Σχήμα 6.5, δηλαδή οι νέες υπό – έννοιες προστίθενται και στο σχήμα της οντολογίας.



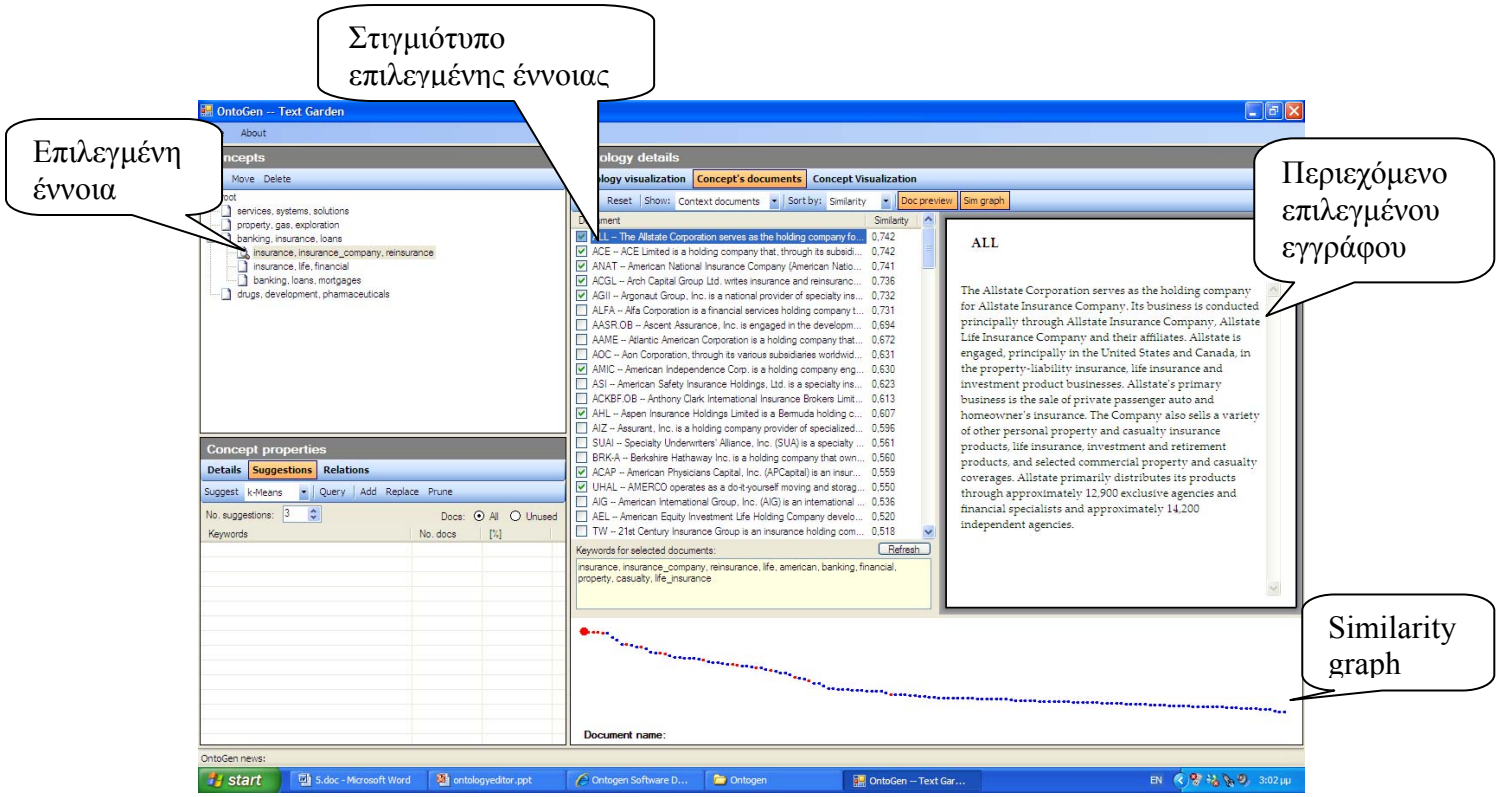
Σχήμα 6.5: Οι νέες υπό – έννοιες της έννοιας που επιλέξαμε προστίθενται στην οντολογία.

Τι μπορούμε να κάνουμε μετά? Ας ελέγξουμε τι υπάρχει μέσα σε μια ομάδα(cluster) που επιλέγουμε από την οντολογία μας. Ότι επιλέγουμε φαίνεται σε κόκκινο χρώμα για να διευκολύνουμε το χρήστη. Έστω λοιπόν επιλέγουμε την υπό-έννοια που φαίνεται στο επόμενο Σχήμα 6.6 :



Σχήμα 6.6: Επιλέγουμε μια υπό – έννοια. Το σύστημα την χρωματίζει με κόκκινο χρώμα.

Προχωρούμε, πατώντας «Concept's documents» για να δούμε αυτό που θέλουμε. Εμφανίζεται το Σχήμα 6.7.



Σχήμα 6.7: Εδώ ο χρήστης ερευνά την επιλεγμένη έννοια, ξεφυλλίζοντας τα έγγραφα της επιλεγμένης έννοιας συμπεριλαμβανομένου και του περιεχομένου τους (πάνω δεξιά) και τη γραφική με τις ομοιότητες των εγγράφων κατά μήκος της έννοιας (κάτω δεξιά).

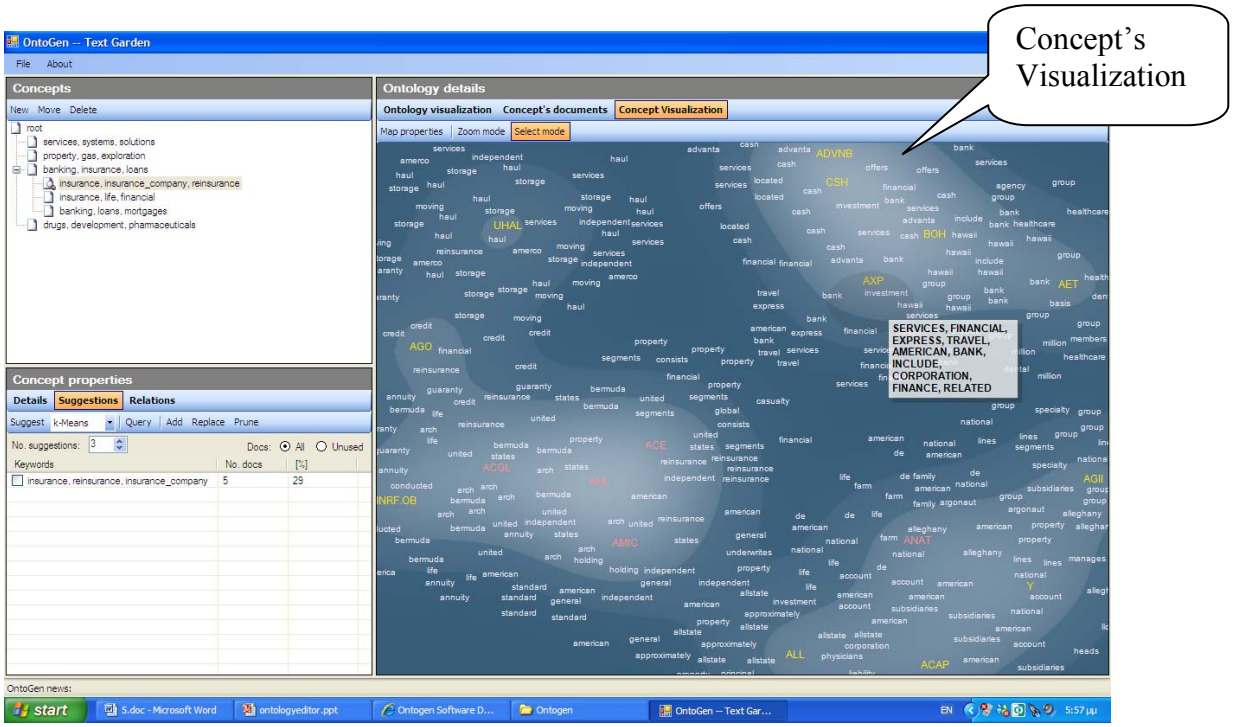
Εδώ βρίσκονται τα κείμενα που εμπίπτουν στη συγκεκριμένη έννοια που επιλέξαμε. Βασικά με το που μια νέα έννοια προστίθεται στην οντολογία, το σύστημα αναθέτει αυτόματα σε αυτήν κάποια στιγμιότυπα. Από τη στιγμή που αυτή η ανάθεση συχνά δεν είναι τέλεια, ο χρήστης και πάλι μπορεί να επέμβει και χειροκίνητα να επαναθέσει στιγμιότυπα από και προς την έννοια. Αυτή η λειτουργικότητα παρέχεται από αυτό το σημείο.

Στα αριστερά είναι μια λίστα με όλα τα στιγμιότυπα και τα στιγμιότυπα που ανήκουν στην έννοια είναι αυτά που είναι επιλεγμένα. Εδώ να διευκρινίσουμε πριν προχωρήσουμε ότι τα στιγμιότυπα είναι τα έγγραφα κειμένων που εμείς φορτώσαμε στο σύστημα σαν είσοδο. Έτσι αν θέλουμε να αλλάξουμε μια ανάθεση ενός στιγμιότυπου για τη συγκεκριμένη έννοια, απλώς δεν το επιλέγουμε και ανάποδα.

Στο κάτω μέρος του Σχήματος 6.7, εμφανίζεται και η γραφική που δίνει την ομοιότητα που υπάρχει μεταξύ των στιγμιότυπων. Η γραφική αυτή είναι πολύ

βοηθητική, αφού μέσω αυτής ο χρήστης μπορεί γρήγορα να προσθέσει ή να αφαιρέσει στιγμιότυπα από τις έννοιες, απλώς επιλέγοντας τα από τη γραφική με το ποντίκι.

Το άλλο που παρέχει το σύστημα είναι, πατώντας το κουμπί «Concept visualization»: προχωρούμε στο Σχήμα 6.8.



Σχήμα 6.8: Αυτή η απεικόνιση δείχνει ένα χάρτη όπου κάθε στιγμιότυπο είναι μια περιγραφή μιας ασφαλιστικής εταιρείας. Μπορούμε να δούμε ότι αυτά τα στιγμιότυπα κατηγοριοποιούνται ομοιόμορφα σε τρείς διαφορετικές ομάδες.

Όπου με αυτό το εργαλείο, μπορούμε να δούμε την οπτική αναπαράσταση των στιγμιότυπων της επιλεγμένης έννοιας.

Η οπτική απεικόνιση είναι ένας χάρτης όπου κάθε στιγμιότυπο είναι μια περιγραφή μιας insurance_company(η επιλεγμένη έννοια). Στο σχήμα βλέπουμε ότι προκύπτουν 3 ομοιόμορφες ξεχωριστές ομάδες.

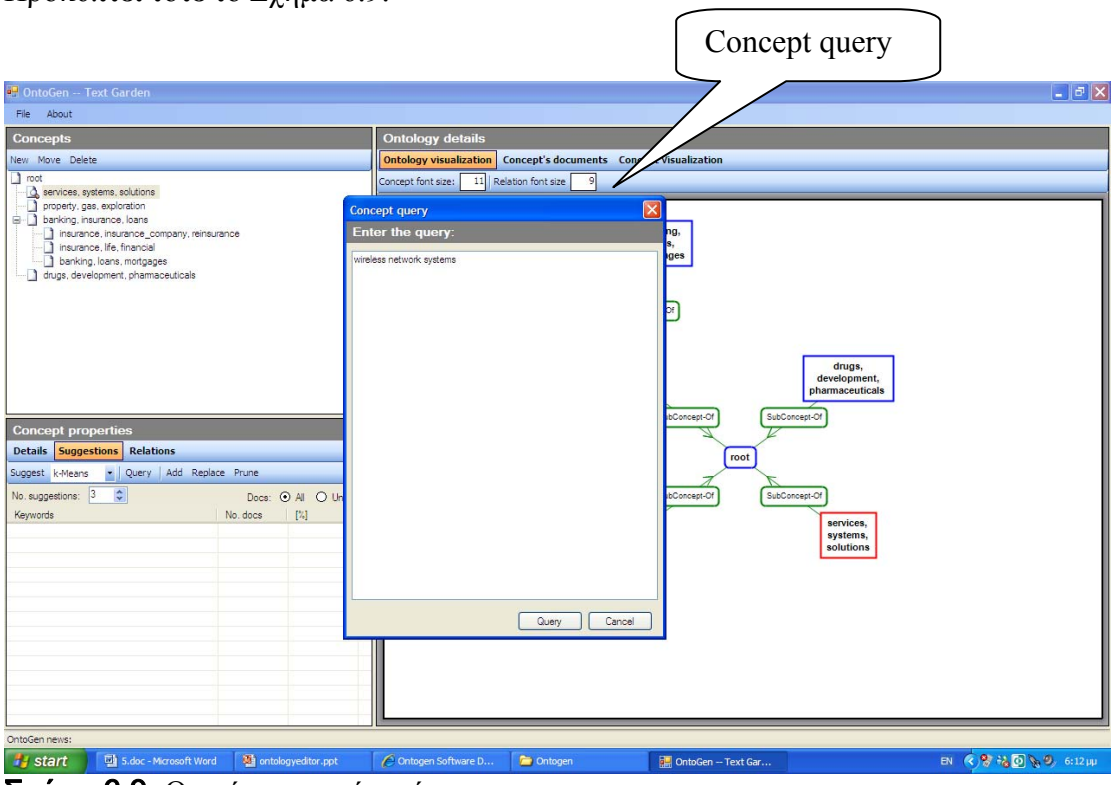
Επίσης η τοποθεσία των ομάδων στο χάρτη έχει σημασία. Δηλ. πόσο κοντά βρίσκεται η μια στην άλλη υποδηλώνει πόσο ομοιότητα υπάρχει στο περιεχόμενο των κειμένων

που περιέχονται σε κάθε ομάδα. Μπορεί να δούμε συνεπώς και 2 ομάδες να έχουν τομή, αν η ομοιότητα που παρουσιάζουν είναι αρκετά μεγάλη ή αντίθετα η μια να είναι αρκετά μακριά από την άλλη αν δεν παρουσιάζουν πολλές ή καθόλου ομοιότητες στο περιεχόμενο.

Ακόμα μπορώ με το ποντίκι να επιλέξω μια περιοχή από το χάρτη και έτσι έχω ακόμα ένα τρόπο για να εισάγω έννοιες στην οντολογία.

Ένας άλλος τρόπος εισαγωγής εννοιών στην οντολογία, είναι βασισμένος στη μέθοδο SVM active learning, και εφαρμόζεται μόνο στα στιγμιότυπα της επιλεγμένης έννοιας.

Συγκεκριμένα, ο χρήστης επιλέγει πάλι την οντολογία στην οποία θέλει να προσθέσει τις νέες υπό – έννοιες και στη συνέχεια από το παράθυρο όπου γίνονται οι Εισηγήσεις πατάει «Query».
Προκύπτει τότε το Σχήμα 6.9.

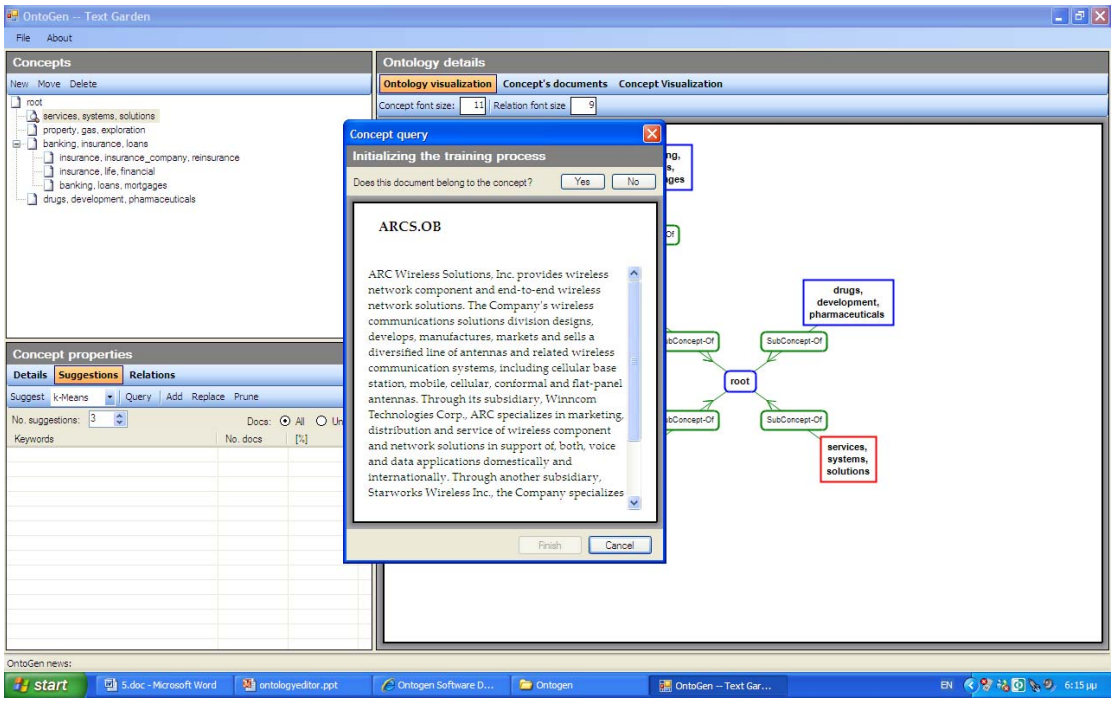


Σχήμα 6.9: Ο χρήστης εισάγει ένα query.

Με κόκκινο χρώμα είναι η επιλεγμένη έννοια. Το παράθυρο που φαίνεται στο Σχήμα 6.9, εμφανίζεται με το πάτημα του κουμπιού «Query», στο οποίο ο χρήστης καλείται

να εισάγει ένα query ή ένα σύνολο λέξεων κλειδιών που να περιγράφει την έννοια που ο χρήστης έχει στο μυαλό του.

Έστω επιλέξαμε από την οντολογία την έννοια: systems, services, solutions και στο παράθυρο για το «Concept Query», εισάγουμε:
Wireless network systems.
Προκύπτει το Σχήμα 6.10.

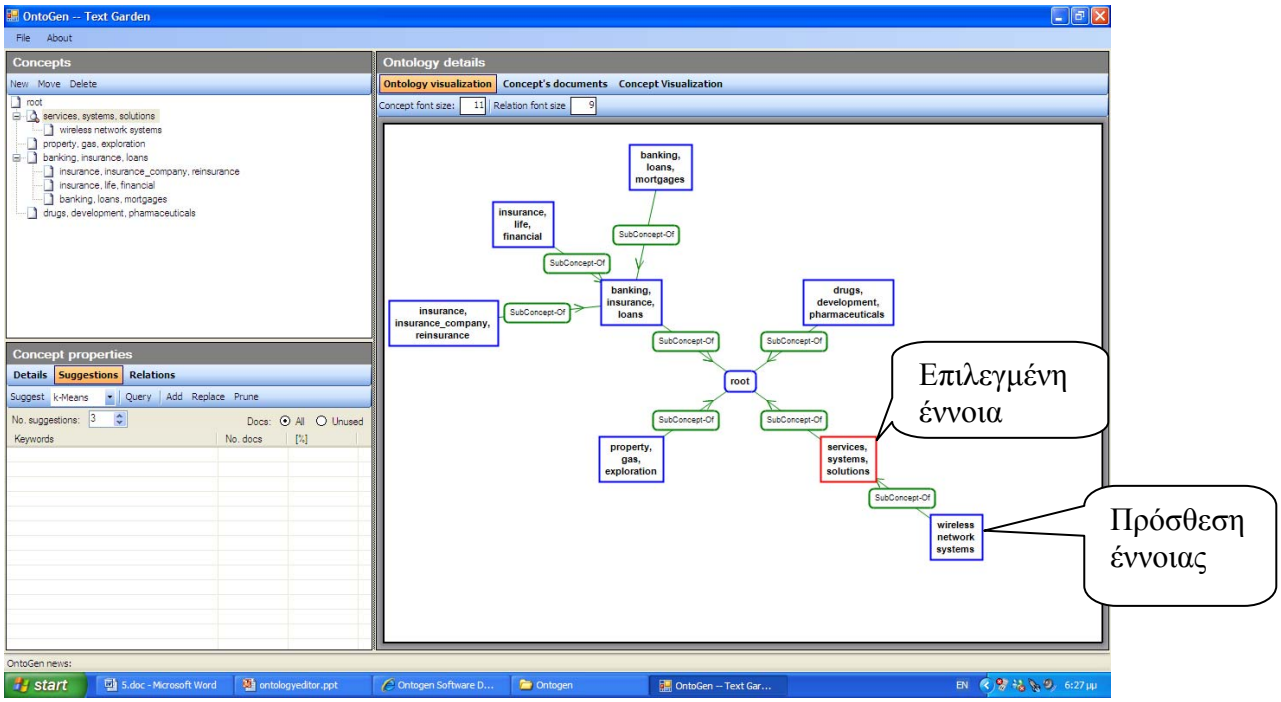


Σχήμα 6.10: Ένα παράδειγμα ερώτησης από το σύστημα για το δοθέν query.

Σε αυτό το σημείο, το σύστημα μας ρωτά, αν το συγκεκριμένο στιγμιότυπο που εμφανίζεται στο Σχήμα (που βασικά είναι ένα κείμενο), ανήκει στην επιλεγμένη έννοια. Ο χρήστης μπορεί να απαντήσει με «Yes» ή «No».

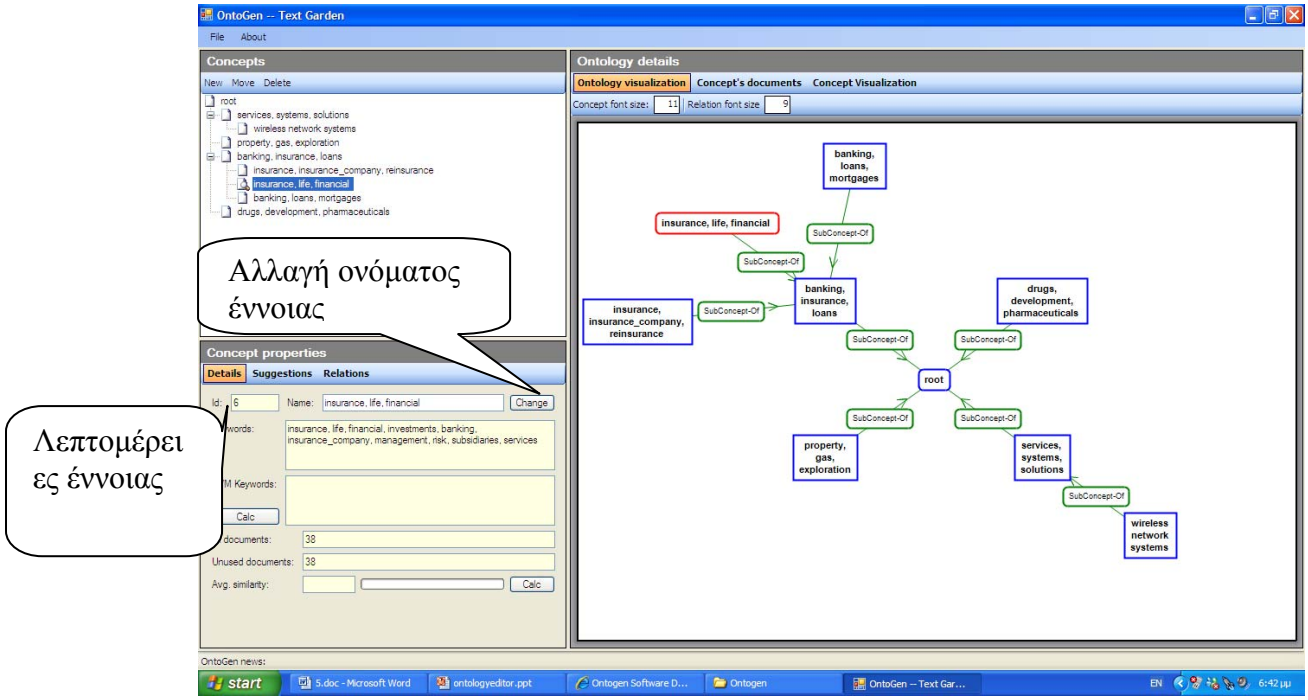
Στο χρήστη εμφανίζονται διάφορα έγγραφα, τα οποία αφήνεται στη δίκη του κρίση να αποφασίσει κατά πόσο ένα στιγμιότυπο ανήκει ή όχι στη συγκεκριμένη έννοια. Αναλόγως με το πόσο ικανοποιημένος είναι ο χρήστης με τις εισηγήσεις, μπορεί να σταματήσει αυτή τη διαδικασία πατώντας «Finish».

Αφού η έννοια κατασκευαστεί, προστίθεται στην οντολογία ως υπό – έννοια της επιλεγμένης έννοιας.



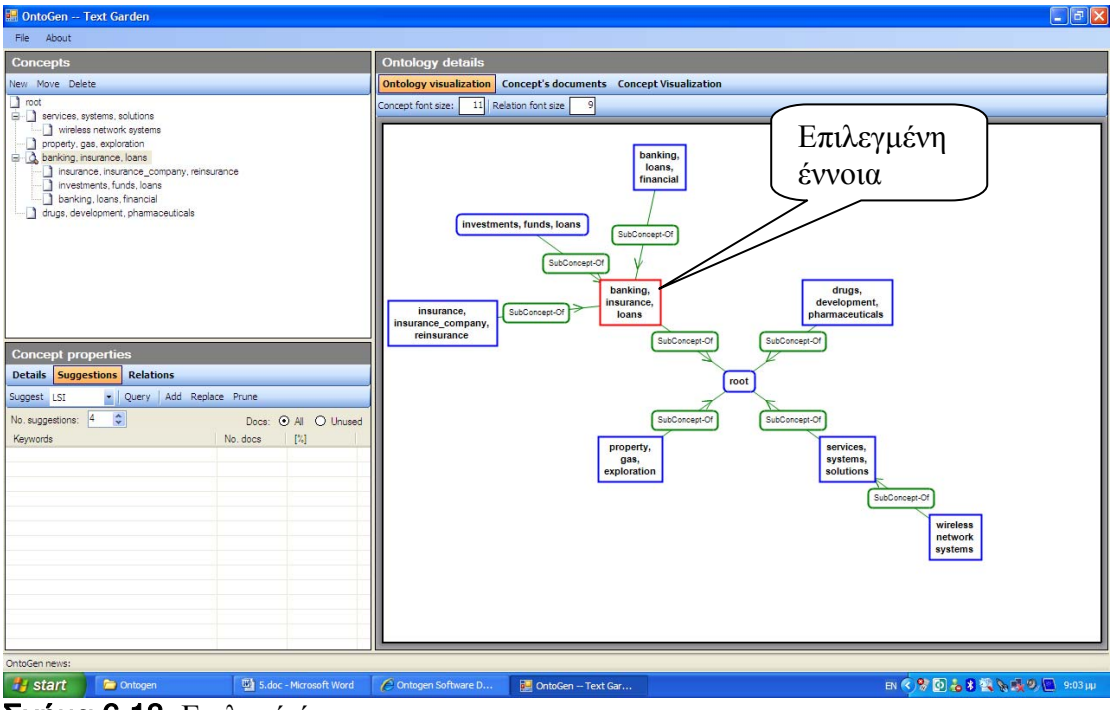
Σχήμα 6.11: Με το τέλος των ερωτήσεων από το σύστημα, η νέα έννοια προστίθεται στην οντολογία.

Επίσης μπορούμε να αλλάξουμε το όνομα μιας έννοιας διαμέσου του παράθυρου «Concept properties». Επιλέγω την έννοια της οποίας θέλω να αλλάξω το όνομα και στο πεδίο «Change» μπορώ να γράψω το νέο όνομα. Πατώντας ακολούθως «Change» η αλλαγή γίνεται.



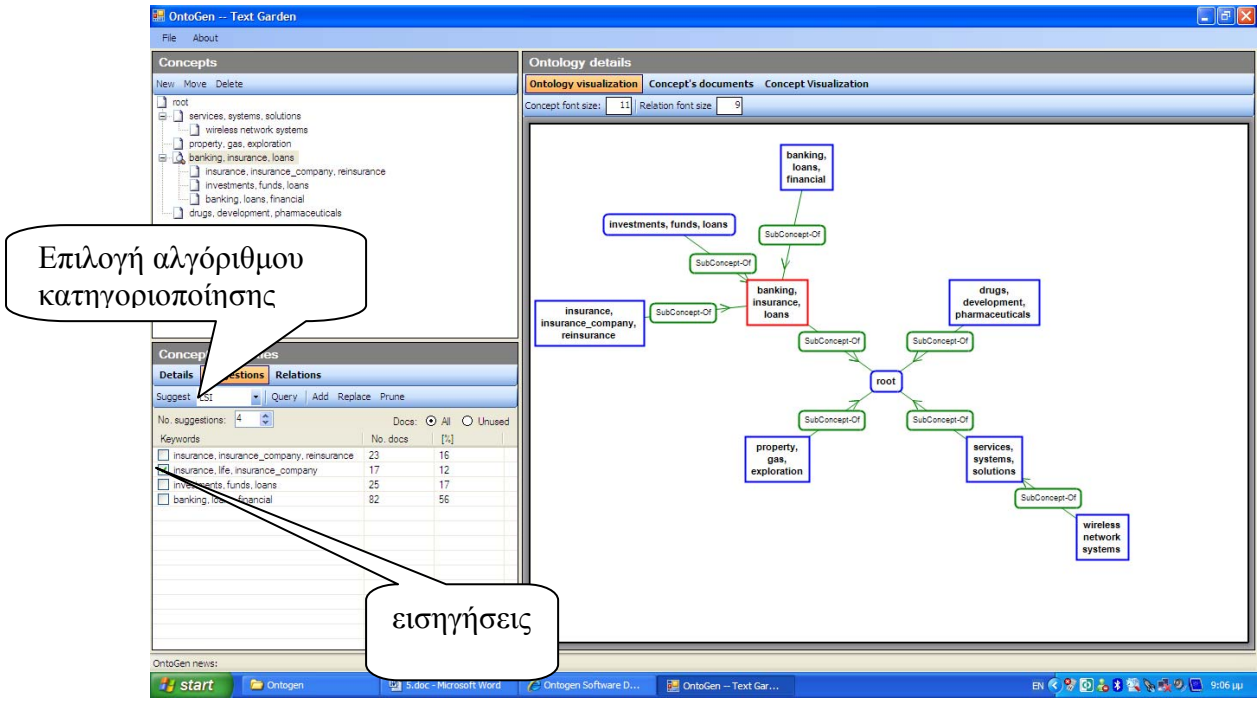
Σχήμα 6.12: Ένας τρόπος αλλαγής των ονομάτων των εννοιών ή υπό – εννοιών στην οντολογία.

Ακόμα μια άλλη λειτουργικότητα που αξίζει να δούμε, αφορά πάλι την πρόσθεση έννοιας ή υπό – εννοιών, αλλά με διαφορετικό τρόπο (ακολουθεί διαφορετικό αλγόριθμο). Η λειτουργικότητα αυτή απεικονίζεται στο Σχήμα 6.13 παρακάτω.



Σχήμα 6.13: Επιλογή έννοιας.

Επιλέγουμε και πάλι την έννοια στην οποία θέλουμε να προσθέσουμε κάποια ή κάποιες υπό – έννοιες.

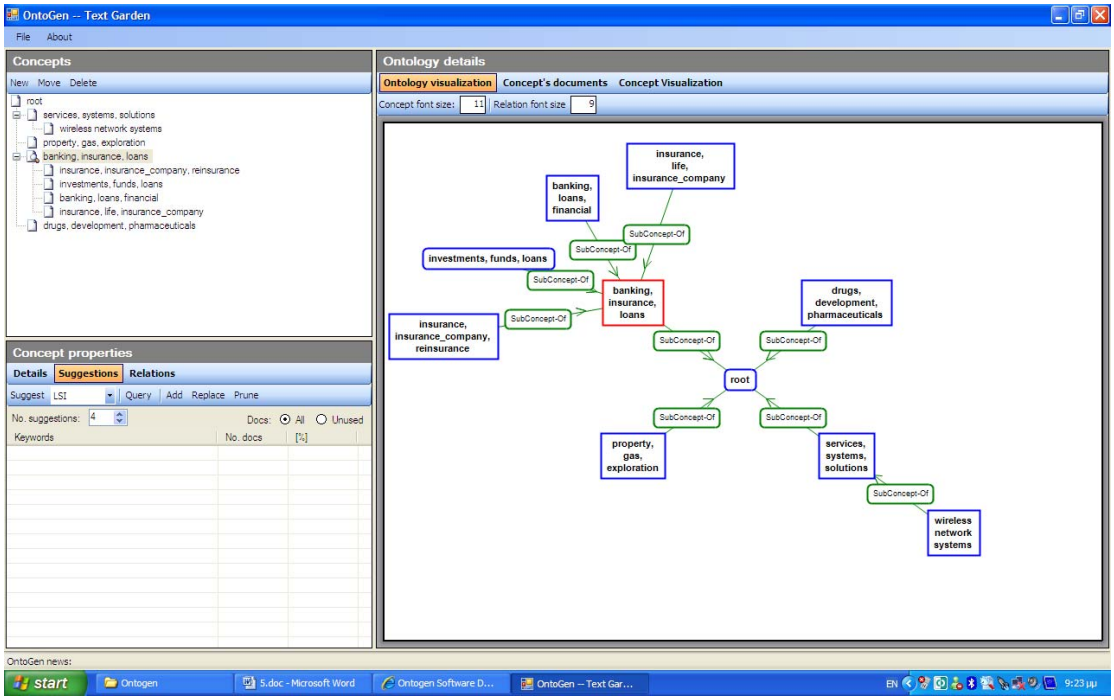


Σχήμα 6.14: Ο χρήστης επιλέγει τον αλγόριθμο κατηγοριοποίησης LSI.

Ακολούθως, όπως βλέπετε και στο Σχήμα 6.14, ο χρήστης καλείται να επιλέξει από το κάτω αριστερό παράθυρο τον αλγόριθμο με τον οποίο προτιμά να γίνει η κατηγοριοποίηση των εγγράφων (LSI ή k-means στο σύστημα). Έπειτα πατώντας «Add», εμφανίζεται η λίστα με τις εισηγήσεις, όπως την είχαμε ξαναδεί πιο πάνω κατά τη διάρκεια της υλοποίησης. Όταν λέμε εισηγήσεις να θυμίσουμε ότι εννοούμε τις προτεινόμενες έννοιες. Συνεπώς, μετά από αυτό το βήμα ο χρήστης πρέπει να επιλέξει ποιες έννοιες πιστεύει πρέπει να προστεθούν ως υπό – έννοιες στην οντολογία για την έννοια που έχουμε επιλέξει.

Στο παράδειγμα μας, επιλέξαμε την έννοια: banking, insurance, loans και μας εμφανίζει μια λίστα τεσσάρων εννοιών. Ο χρήστης επιλέγει μόνο την μία εκ των τεσσάρων διότι οι άλλες τρείς ήδη είναι υπό – έννοιες της συγκεκριμένης έννοιας, τις οποίες είχαμε προσθέσει προηγουμένως.

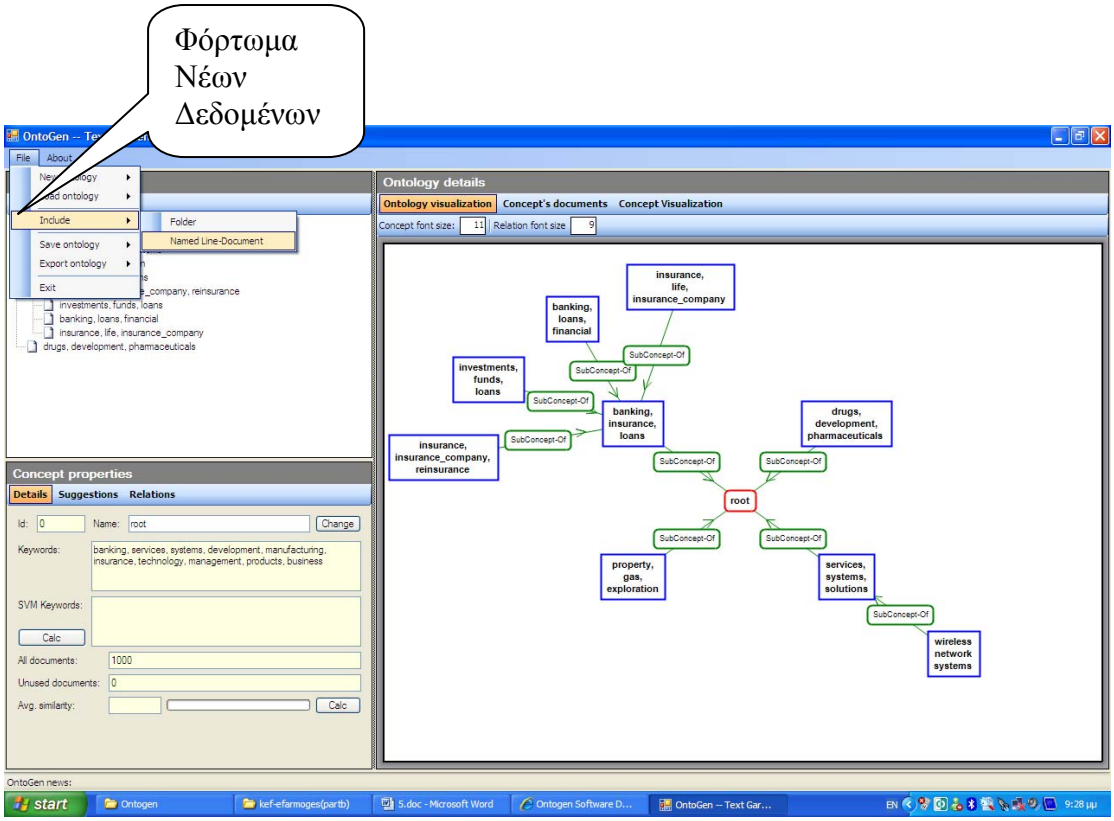
Με το «Add» λοιπόν προκύπτει το παρακάτω Σχήμα 6.15:



Σχήμα 6.15: Η μορφή της οντολογίας μας στο παρόν στάδιο, μετά τη πρόσθεση των νέων υπό – εννοιών με τον αλγόριθμο LSI.

Τι γίνεται όμως όταν έχω καινούργια έγγραφα – κείμενα και θέλω να τα προσθέσω σε αυτή την οντολογία.

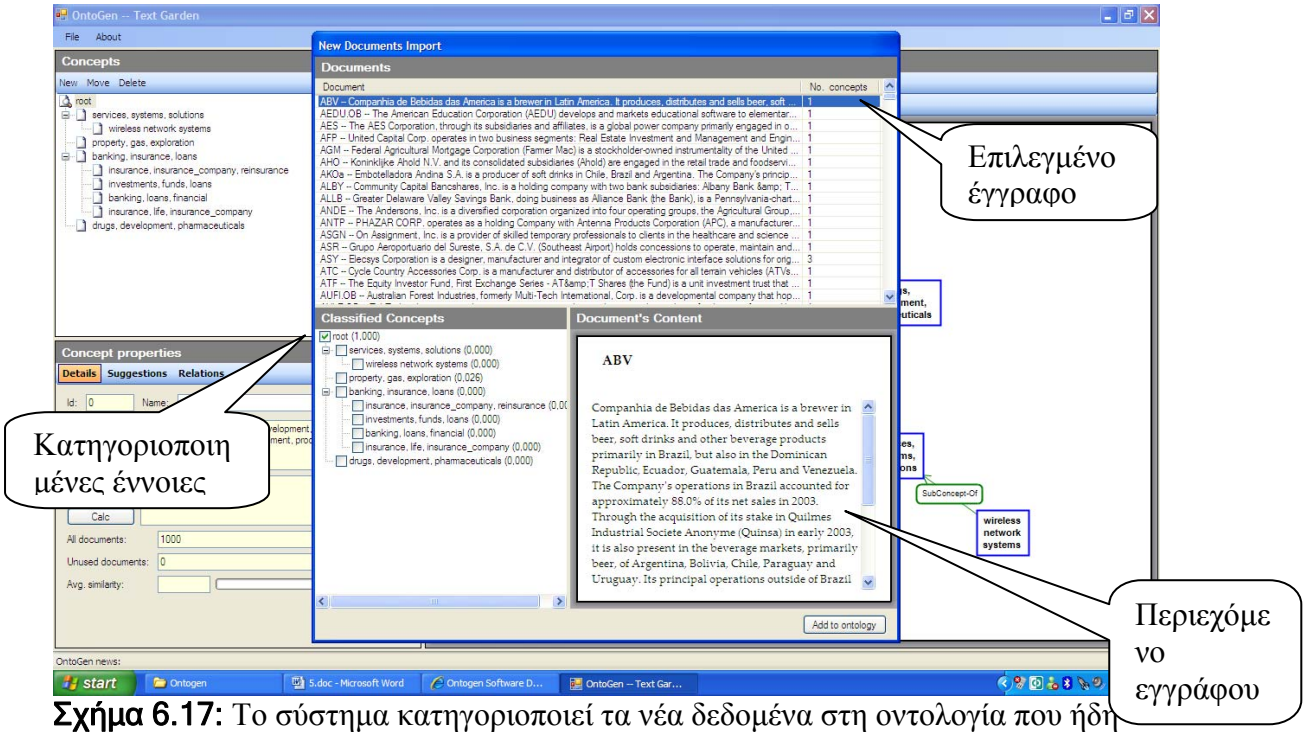
Τότε παρέχεται και αυτή η λειτουργικότητα από το σύστημα, δηλαδή να μπορώ να συμπεριλάβω νέα δεδομένα σε ήδη μια υπάρχουσα οντολογία που βρίσκεται σε εξέλιξη και αυτό όπως απεικονίζεται και στο Σχήμα 6.16, μπορεί να γίνει διάμεσο του File>Include>Named Line Document.
Μετά εμφανίζεται ένα παράθυρο από το οποίο κάμνοντας browse, βρίσκουμε τη νέα βάση με τα αρχεία που θέλουμε να συμπεριλάβουμε.



Σχήμα 6.16: Ενημέρωση μιας ήδη υπάρχουσας οντολογίας με νέα δεδομένα.

Αφού τα νέα δεδομένα φορτωθούν στο σύστημα, μεσολαβεί ένα πολύ μικρό χρονικό διάστημα για να γίνει βασικά αυτή η φόρτωση αλλά και για να εκπαιδεύσει το σύστημα το ταξινομητή ώστε να κατηγοριοποιήσει τα δεδομένα στην ήδη υπάρχουσα οντολογία.

Και παρακάτω στο Σχήμα 6.17 έχουμε τα αποτελέσματα.



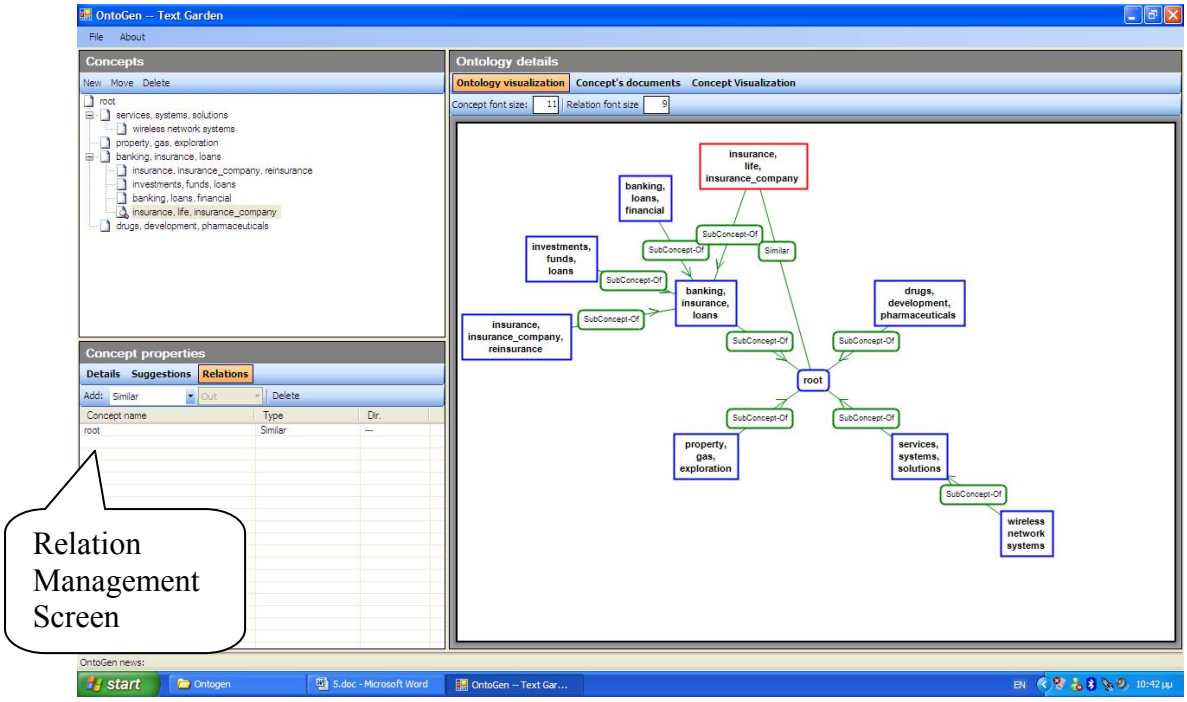
Σχήμα 6.17: Το σύστημα κατηγοριοποιεί τα νέα δεδομένα στη οντολογία που ήδη είναι δημιουργημένη στο σύστημα.

Και όταν λέμε αποτελέσματα της κατηγοριοποίησης των εγγράφων, για παράδειγμα το πρώτο έγγραφο που είναι επιλεγμένο (με μπλε χρώμα στο Σχήμα 6.17), αν το επιλέξεις στο κάτω αριστερό παράθυρο σου εμφανίζει σε ποια έννοια το κατηγοριοποίησε το σύστημα. Το συγκεκριμένο έγγραφο κατηγοριοποιείται στη root concept.

Μπορεί όμως και ο ίδιος ο χρήστης να επέμβει στην κατηγοριοποίηση. Πιο συγκεκριμένα μπορεί επιλέγοντας ένα κείμενο να επιλέξει για αυτό σε ποια έννοια να κατηγοριοποιηθεί και έπειτα αφού τελειώσει με όλες τις αλλαγές που επιθυμεί να κάμει, πατάει «Add to ontology».

Επιπρόσθετα, όσον αφορά το κομμάτι των Σχέσεων (Relations) μεταξύ των εννοιών και υπό – εννοιών κυρίως στο σύστημα, ο χρήστης μπορεί να δημιουργήσει και άλλους τύπους σχέσεων εκτός της σχέσης «sub-concept» που ήδη υπάρχει. Αυτό παρέχεται διαμέσου του Relation management στο αριστερό κάτω μέρος του Σχήματος, όπου εκεί ο χρήστης μπορεί να δει μια λίστα με όλες τις σχέσεις της

επιλεγμένης έννοιας. Μπορεί να προσθέσει νέα σχέση, να διαγράψει ήδη υπάρχουσες ή να προσθέσει νέους τύπους σχέσεων (Σχήμα 6.18).



Σχήμα 6.18: Δημιουργία νέων τύπων σχέσεων μεταξύ εννοιών ή υπό – εννοιών.

6.4 Αξιολόγηση Συστήματος

Τα εργαλεία παραγωγής Οντολογιών διαφέρουν το ένα από το άλλο σε πολλές διαστάσεις συμπεριλαμβανομένου των λειτουργιών που διαθέτουν, των επιπέδων στήριξης που παρέχουν στο χρήστη, την ικανότητα που έχουν να χειρίζονται δεδομένα, δυνατότητες οπτικής αναπαράστασης και κάποια άλλα ίσως.

Με σκοπό να αξιολογηθούν τα αποτελέσματα της παραγωγής οντολογίας, για παράδειγμα την οντολογία την ίδια, είναι δυνατόν να αξιολογηθεί το εργαλείο από τη πλευρά των χρηστών. Γενικά, εάν μια νέα τεχνολογία, εργαλείο ή προϊόν εναρμονίζεται πλήρως με τη καθημερινή ζωή και γίνεται μέρος της καθημερινής ζωής δηλαδή, τότε πρέπει να γίνεται αποδεκτό από τους χρήστες (Barnum, 2001).

Έτσι λοιπόν για να διεξαχθεί αυτή η μελέτη χρηστών και να αξιολογηθεί το λογισμικό εργαλείο OntoGen για την παραγωγή οντολογίας, εφαρμόστηκαν δοκιμές χρηστών πάνω σε δύο ομάδες χρηστών. Μέσα από αυτές τις δοκιμές καταγράφονται οι αντιδράσεις των χρηστών.

Σκεπτόμενοι οι αξιολογητές του συστήματος ποιες θα είναι οι δύο ομάδες των χρηστών όρισαν πρώτα τη δομή του δείγματος του χρήστη, το δείγμα για το προφίλ, τον αριθμό και άλλα χαρακτηριστικά του συγκεκριμένου κοινού στο οποίο θα απευθυνθούν(αξιολογητές). Η μεθοδολογία περιείχε προδιαγραφές του εργαλείου υπό αξιολόγηση, οι οποίες είχαν μια επιρροή στα αποτελέσματα και τις μεθόδους οι οποίες χρησιμοποιήθηκαν στις δοκιμές των χρηστών.

Η μεθοδολογία περιελάμβανε δύο ερωτηματολόγια τα οποία συμπληρώθηκαν από τον καθένα από τους χρήστες που συμμετείχαν στη μελέτη αυτή. Τα δύο αυτά ερωτηματολόγια ήταν: ένα Αρχικό Ερωτηματολόγιο και το Τελικό Ερωτηματολόγιο. Ακολουθούσε το Standard protocol, το οποίο περιελάμβανε δοκιμαστική φάση και ανάλυση δεδομένων.

6.4.1 Σκοπός της αξιολόγησης

Η λογική πίσω από τη διεξαγωγή μιας μελέτης χρηστών είναι ότι θα αυξήσει την πιθανότητα ότι το εργαλείο, η υπηρεσία ή τα τελικά προϊόντα τα οποία αξιολογούμε θα ικανοποιήσουν τις ανάγκες του χρήστη. Με την έναρξη μιας καινοτόμους διαδικασίας, οι ανάγκες του χρήστη και οι γενικές προσδοκίες δεν μπορούν να προσδιοριστούν καλά, εξελίσσονται ως ευκαιρίες που ξυπνούν μέσα από την προσπάθεια τους να δοκιμάσουν νέα προϊόντα και τεχνολογίες και να τα συσχετίσουν με την εργασία τους.
Έτσι λοιπόν οι χρήστες είναι μια πολύτιμη πηγή δεδομένων και ανατροφοδότησης οι οποίοι μπορούν να εμπλουτίσουν την ποιότητα του εργαλείου που έχει αναπτυχθεί μέχρις στιγμής.

Στη δική μας μελέτη χρηστών, αυτοί που διεξήγαγαν την όλη έρευνα, κυρίως έδειξαν ενδιαφέρον για τη γνώμη του νεότερου πληθυσμού με συγκεκριμένα ενδιαφέροντα στην Επιστήμη Υπολογιστών και στις Γνωστικές Επιστήμες (Ψυχολογία και Παιδαγωγικά).

Αποφάσισαν να χρησιμοποιήσουν τη δοκιμή αξιολόγησης, έτσι ώστε να ελέγξουν την αποτελεσματικότητα του εργαλείου, από πλευράς σχεδίασης, λειτουργικότητας και διεπαφής. Κατά συνέπεια, οι συγκεκριμένες ανεπάρκειες μη υλοποίησης κάποιων στόχων και οι αιτίες που τις προκαλούν μπορούν να προσδιοριστούν. Ο στόχος της μελέτης αυτής ήταν να εξεταστεί κατά πόσο χρησιμοποιείται το σύστημα και είναι χρήσιμο ούτως ώστε ενδεχομένως να βελτιώσει την εμπειρία χρήσης για τον καθένα που χρησιμοποιεί την πλατφόρμα.

6.4.2 Ομάδες Χρηστών

Διαφορετικοί χρήστες έχουν διαφορετικές ιδέες και μορφές οργάνωσης της γνώσης. Μερικές φορές αυτές οι διαδικασίες είναι υποσυνείδητα αισθητές και οι άνθρωποι δεν είναι ενήμεροι αυτών. Τα εργαλεία λογισμικού για την κατασκευή οντολογίας υποστηρίζουν διαφορετικές δραστηριότητες συμπεριλαμβανομένου την παραγωγή και διαχείριση των δικτύων των εννοιών γνώσης, το οποίο είναι γενικά απαιτητικό να κατανοηθεί και απαιτεί χρόνο και συμμετοχή για να επιτευχθεί.

Ωστόσο, επέλεξαν δύο ομάδες χρηστών για να συμμετέχουν στη μελέτη των χρηστών για το εργαλείο OntoGen για την κατασκευή οντολογιών.

Την πρώτη ομάδα αποτελούσουν φοιτητές πληροφορικής, ως τελικοί χρήστες του συστήματος με υψηλές τεχνικές γνώσεις αλλά με μέσο επίπεδο όσον αφορά τις γνωστικές τους γνώσεις. Τη δεύτερη ομάδα αποτελούσαν φοιτητές του τμήματος ψυχολογίας και των παιδαγωγικών, ως τελικοί χρήστες και αυτοί βεβαίως, έχοντας όμως τώρα αυτοί μέσο επίπεδο γνώσεων σε τεχνικά θέματα και υψηλές γνώσεις σε γνωστικές διαδικασίες. Η ομάδα των χρηστών διαμορφώθηκε με τη λήψη όλων των φοιτητών που συμμετείχαν εκείνη την περίοδο στο κανονικό πρόγραμμα στη Σχολή Τεχνών και Επιστημών του Πανεπιστημίου του Rijeka.

Η μελέτη των χρηστών πραγματοποιήθηκε στη Σχολή Τεχνών και Επιστημών, του Πανεπιστημίου του Rijeka στην Κροατία. Η έρευνα περιέλαβε 91 φοιτητές, 48 (24 γυναίκες, 24 άνδρες) από το Τμήμα Πληροφορικής, 34 (29 γυναίκες, 5 άνδρες) από το Τμήμα Ψυχολογίας και 9 (όλες γυναίκες) από το Παιδαγωγικό Τμήμα. Το διάστημα ηλικίας των χρηστών που συμμετείχαν κυμαινόταν από 19-29 χρονών (με μέσο όρο 20.94 και μέση απόκλιση 1.41).

6.4.3 Μεθοδολογία Αξιολόγησης

Όπως αναφέρθηκε και εισαγωγικά η μεθοδολογία της έρευνας που έγινε εμπεριέχει δύο ερωτηματολόγια τα οποία έχουν αναπτυχθεί για τους σκοπούς της έρευνας. Το ένα χρησιμοποιήθηκε για να εκτιμηθούν οι προτιμήσεις του χρήστη και οι δεξιότητες χρήσης υπολογιστών, όπως επίσης και τη γνώση τους σχετικά με τις γνωστικές διαδικασίες (Αρχικό ερωτηματολόγιο). Περιέχει 15 ερωτήσεις. Συμπληρώθηκε από τους χρήστες πριν από την αλληλεπίδραση τους με το εργαλείο. Το δεύτερο ερωτηματολόγιο σχεδιάστηκε για να αξιολογηθεί η εμπειρία των χρηστών με το εργαλείο και δόθηκε στους χρήστες για συμπλήρωση μετά τη δοκιμή (Τελικό ερωτηματολόγιο). Περιέχει 24 ερωτήσεις.

Το **αρχικό ερωτηματολόγιο** ήταν το πρώτο βήμα στη μελέτη χρηστών. Το ερωτηματολόγιο οργανώθηκε σε διάφορα θέματα καλύπτοντας έτσι τη χρήση του υπολογιστή, της συνήθειες μάθησης και την οργάνωση γνώσης.

Η **χρήση του ηλεκτρονικού υπολογιστή** ελέγχτηκε με διάφορες ερωτήσεις στην περιοχή της εμπειρίας των υπολογιστών και ικανοτήτων (με κλίμακα 0-4). Αυτό περιελάμβανε ικανότητες χειρισμού υπολογιστή και χρόνος που καταναλώνεται στον υπολογιστή σε διάφορες περιοχές, όπως:
Email, Chat, Web surfing, Web searching, e-learning system, Windows, Linux and Mac.

Συνήθειες Εκμάθησης. Το μέρος αυτό αξιολόγησε τις προτιμήσεις εκμάθησης τους, όπως: να μαθαίνουν ομαδικά, μόνοι, διαβάζοντας, μιλώντας σε άλλους, ακούοντας,

είτε μέσω εξάσκησης. Επίσης περιελήφθησαν ερωτήσεις που συλλαμβάνουν την άποψη σχετικά με τη χρησιμοποίηση τεχνολογίας κατά τη διάρκεια της μάθησης.

Οργάνωση γνώσης. Το μέρος αυτό του ερωτηματολογίου καλύπτει τις διανοητικές διαδικασίες τις οποίες αναγνωρίζει κατά τη διάρκεια της εκμάθησης και οργάνωσης της γνώσης, όπως επίσης και την επίσημε γνώση τους για αυτή την περιοχή (π.χ. cognitive maps). Περιέλαβε επίσης ποιοτικές ερωτήσεις σχετικά με τους όρους «γνωστικοί χάρτες», με τις μεθόδους εκμάθησης και οργάνωσης της γνώσης, και την ιδέα για το πιθανό υπολογιστικό εργαλείο το οποίο θα διευκόλυνε την οργάνωση γνώσης.

Το **Τελικό ερωτηματολόγιο** που δόθηκε στους χρήστες στο τέλος της δοκιμαστικής φάσης, χρησιμοποιήθηκε για να ληφθούν τα τελικά συμπεράσματα από το εργαλείο που αξιολογείται.
Οι ερωτήσεις που περιελήφθησαν ήταν σχετικές με:

- Γενική εντύπωση του εργαλείου
- Η αξιολόγηση του layout των κύριων λειτουργιών και η γραφική διεπαφή τους
- Έννοια και διαχείριση εγγράφων

Τα ερωτήματα είχαν κλίμακα βαθμολόγησης 1-5 με εξήγηση: 1- κακό, 2- ικανοποιητικό, 3- καλό, 4- πολύ καλό, 5- άριστο.

Η εμπειρία τους στη δημιουργία και τη διάσωση οντολογιών εξετάστηκαν ποιοτικά, καθώς επίσης και οι μέθοδοι εισηγήσεων εννοιών και οι οπτικές απεικονίσεις οντολογίας.

6.4.4 Διαδικασίες και Περιγραφή δεδομένων

Οι διαδικασίες αντιπροσωπεύουν το ακριβές πρωτόκολλο που ακολουθήθηκαν σε κάθε περίοδο της διευθετημένης μελέτης χρηστών.

Στην περίπτωση μας χωρίζονται σε τρείς φάσεις:

- **Εισαγωγή** – η οποία παρουσιάζει τους κύριους στόχους της μελέτης
- **Έλεγχος και Συλλογή Δεδομένων** – εδώ γίνεται η παρουσίαση και το συμπλήρωμα του Αρχικού ερωτηματολογίου, επίδειξη του εργαλείου που θα αξιολογηθεί, καθώς επίσης και παρουσίαση και συμπλήρωμα του Τελικού ερωτηματολογίου
- **Συμπέρασμα** – διεξάγεται μια συζήτηση με βάση την εμπειρία που απέκτησαν οι χρήστες κατά τη διάρκεια επαφής τους με το εργαλείο.

Οι καθοδηγημένες εργασίες, οι οποίες περιλάμβαναν χειρισμό σε δύο σύνολα δεδομένων, ένα με προφίλ επιχειρήσεων παρμένα από το Yahoo!Web site και το άλλο περιείχε άρθρα ειδήσεων.

Τα δεδομένα σχετικά με τις επιχειρήσεις περιλάμβαναν προφίλ 7177 εταιρειών. Τα δεδομένα χωρίστηκαν σε δύο μέρη: με το κύριο μέρος να έχει 6177 εταιρείες και το υπόλοιπο μέρος να έχει 1000 εταιρείες, το οποίο χρησιμοποιήθηκε για τον έλεγχο της ικανότητας του OntoGen να μαθαίνει ενεργά (active learning). Ο στόχος με αυτό το σύνολο δεδομένων ήταν να κατασκευαστεί η οντολογία η οποία συλλαμβάνει τις περιοχές που καλύπτονται από τις επιχειρήσεις μέσα στη συλλογή.

Το άλλο σύνολο δεδομένων, με τα άρθρα ειδήσεων, περιελάμβανε ένα δείγμα των 500 ειδησεογραφικών άρθρων από το πρακτορείο ειδήσεων Reuters. Τα δεδομένα διαχειρίστηκαν με δύο τρόπους: από τις χώρες στις οποίες εμφανίστηκαν αυτά και κατά θέμα. Έτσι, στόχος ήταν να κατασκευαστούν τώρα δύο διαφορετικές οντολογίες με βάση τα ίδια δεδομένα. Στην πρώτη περίπτωση η άποψη ήταν γεωγραφική και στη δεύτερη περίπτωση ήταν βασισμένη στο θέμα.

6.4.5 Αποτελέσματα και Εισηγήσεις

Η ανάλυση του Αρχικού ερωτηματολόγιου έδειξε σημαντικές στατιστικές διαφορές στη συχνότητα χρήσης σε κάποια από τα προγράμματα και υπηρεσίες (email, chat

and surfing the web), αλλά δεν υπήρχε σημαντική διαφορά στη χρήση web searching, e-learning system, Linux και Mac OSX (βλέπετε **Πίνακας 8.1**) . Η σημαντική όμως διαφορά υπήρξε στη τυπική γνώση των διαδικασιών μάθησης, όπως αναμενόταν, οι περισσότεροι φοιτητές της Ψυχολογίας και του Παιδαγωγικού ήξεραν ακριβώς τι ήταν οι "Γνωστικοί Χάρτες". Με το ίδιο σκεπτικό, υπήρχε σημαντική διαφορά στη χρήση υπολογιστών από την πλευρά των φοιτητών Πληροφορικής όπως επίσης και στο ποσοστό του μέσου όρου που καταναλώνουν δουλεύοντας στους υπολογιστές τους. Δεν υπήρχε διαφορά στη χρήση Linux και Mac OSX επειδή φοιτητές από όλα τα συμμετέχοντα Τμήματα συνήθως δεν τα χρησιμοποιούν αυτά τα λειτουργικά συστήματα. Επίσης δεν υπήρχε διαφορά στο χρόνο που καταναλώνεται για searching στο Web.

Η ανάλυση του Τελικού ερωτηματολογίου αποκάλυψε ότι γενικά οι χρήστες έχουν μια καλή εντύπωση για το εργαλείο OntoGen (75% είπαν «καλό» με «πολύ καλό»). Οι περισσότεροι από τους χρήστες πιστεύουν ότι το εργαλείο αυτό είναι εύχρηστο (93%), και γύρω στους μισούς από αυτούς αναφέρονται στο σχεδιάγραμμα των τριών κύριων μερών της διεπιφάνειας, σαν «κυρίως καλό» (Ontology visualisation -44%, Concept tree - 45%, Concept Properties – 50%). Παρόμοια κατάσταση συνέβηκε στην αξιολόγηση της γραφικής διεπιφάνειας. Η γενική εντύπωση είναι και πάλι «κυρίως καλή», αλλά κάποια αντικείμενα αξιολογήθηκαν λαμβάνοντας υπόψη πόσο ελκύουν το χρήστη (36% δεν τη βρήκα ιδιαίτερα ελκυστική ή καθόλου ελκυστική). Οι χρήστες συνέχεια τόνιζα το γεγονός ότι θα έπρεπε να υπήρχαν περισσότερα χρώματα.

Οι ποιοτικές αναλύσεις έδειξαν ότι ένα από τα κύρια πλεονεκτήματα του OntoGen είναι ο χειρισμός μεγάλων βάσεων δεδομένων με ευκολία. Σύμφωνα με τους χρήστες, το εργαλείο είναι αποδοτικό, κερδίζει χρόνο και προσπάθεια και δίνει πολύ χώρο στο χρήστη έτσι ώστε να μπορεί να επεμβαίνει. Μειονεκτήματα, έτσι όπως παρατηρήθηκαν, κυρίως αφορούν τη μη ελκυστική όψη, αφαιρετική σύλληψη, περιστασιακή βραδύτητα και η ανάγκη να το μάθεις πως το χρησιμοποιείς πρώτα.
Μετά που έχει γίνει μια δοκιμή του συστήματος για μόνο δύο ώρες, το 75% των χρηστών μπορούσαν εύκολα να ξεχωρίσουν τη διαφορά μεταξύ των δύο τρόπων δημιουργίας οντολογιών, δηλαδή της μη επιβλεπόμενης μεθόδου μάθησης (clustering) και της επιβλεπόμενης μεθόδου μάθησης (active learning).

Η ελλιπής οδηγίες και η ανάγκη για όλες τις δυνατότητες επιλογής σε ένα μέρος, ερμηνεύτηκαν ως οι κύριες δυσκολίες για την παραγωγή νέων εννοιών. Επίσης χαρακτήρισαν τη βασική διαχείριση έννοιας ως «κυρίως» έως «πολύ εύκολη» με αξιόλογα σχόλια.

Η οπτική απεικόνιση των εννοιών αποδείχτηκε να αποτελεί μεγάλη βοήθεια στην επιλογή των υπό – εννοιών, για την πλειοψηφία των χρηστών (90%). Οι περισσότεροι των χρηστών βρήκαν την οπτική απεικόνιση των εννοιών κατάλληλη, αλλά πρόσθεσαν ότι χρειάζεται περισσότερες ελκυστικές γραφικές λύσεις. Επιπλέον, οι χρήστες σκέφτηκαν ότι με την πρόσθεση εννοιών, οι όλη απεικόνιση των εννοιών γίνεται ακατάστατη γεγονός που δυσκολεύει τη μελέτη του χάρτη. Επίσης πιστεύουν ότι το μέρος που αφορά την ενεργή μάθηση μπορεί να βελτιωθεί (αλλά δεν είναι σίγουροι πως) και προτείνουν ακόμη να γίνει επέκταση του εργαλείου αυτού με το να μπορεί ο χρήστης να εισάγει χειροκίνητα νέα έγγραφα.

Η ποιοτική ανάλυση για την οπτική αναπαράσταση των οντολογιών, έδειξε ότι σε αυτό το επίπεδο που βρίσκεται είναι ικανοποιητική, αλλά θα μπορούσε να περιλαμβάνει ψάξιμο (search) με ένα απλό κλικ του ποντικιού.

Όσον αφορά το διαχειρισμό των εγγράφων της έννοιας, ο χρήστης πιστεύει ότι είναι αρκετά καλός. Δεν υπάρχουν σημαντικές διαφορές στην αξιολόγηση αυτού του κομματιού του OntoGen. Μια γενική αξιολόγηση για όλες τις λειτουργίες που περιλαμβάνονται στις λειτουργίες της διαχείρισης των εγγράφων είναι «πολύ καλή».

Οι μικρότερες αντιρρήσεις αφορούσαν τη διαχείριση των σχέσεων, όπου παίρνουμε εισηγήσεις για να συμπεριλάβουμε κάποιο απλό χειρισμό στα έγγραφα. Υπήρξε ένα γενικό σχόλιο ότι η παρουσίαση μιας ομοιότητας ενός εγγράφου σε μια έννοια πρέπει να απλουστευθεί.

Sample N=91	1 Rarely	2 Oftenly	SIGNIFICANCE
	E-MAIL		
Psychology/ Pedagogy	11	32	**Fisher's Exact Test***
Computer sciences	3	45	**6,511; p= 0,018**
	CHAT		
Psychology/ Pedagogy	40	3	
Computer sciences	33	15	**8,422; p = 0,04**
	SURFING THE WEB		
Psychology/ Pedagogy	17	26	
Computer sciences	3	45	**14,654; p = 0,00**
	SEARCHING THE WEB		
Psychology/ Pedagogy	7	36	
Computer sciences	7	41	0,050; p = 1,000
	e-LEARNING SYSTEM		
Psychology/ Pedagogy	43	0	
Computer sciences	44	4	3,748; p = 0,119
	WINDOWS		
Psychology/ Pedagogy	4	39	
Computer sciences	1	47	2,276; p = 0,185
	LINUX		
Psychology/ Pedagogy	43	0	
Computer sciences	45	3	2,779; p = 0,244
	MAC OSX		
Psychology/ Pedagogy	43	0	
Computer sciences	47	1	0,906; p = 1,000

Πίνακας 8.1 : Σύγκριση της συχνοτήτων μεταξύ των δύο ομάδων χρηστών για διαφορετικές υπηρεσίες και προγράμματα: e-mail, chat, surfing the Web, searching the Web, e-learning system, Windows, Linux, Mac. Για κάθε ομάδα δόθηκε το επίπεδο σημαντικότητας. Οι πρώτες τρείς υπηρεσίες (-mail, chat, surfing the Web) έδειξαν σημαντική διαφορά.
(* η ανάλυση εκτελέστηκε με SPSS)

6.5 Συμπεράσματα

Μέσα από μια σειρά βημάτων έχουμε δείξει πως κατασκευάζεται μια οντολογία από ένα μεγάλο αριθμό κειμένων.

Το σημαντικότερο όσο αφορά τη λειτουργία του συστήματος και πρέπει να τονιστεί ακόμα μια φορά είναι ότι το σύστημα ναι μεν αυτοματοποιεί πολλές λειτουργίες και απλοποιεί πολλές από αυτές οι οποίες παλαιότερα ήταν χρονοβόρες και δύσκολες, εν τούτοις η τελική απόφαση σε σχεδόν όλες τις υπό – λειτουργίες του συστήματος υπόκεινται στην κρίση του χρήστη. Το OntoGen δεν είναι ένα σύστημα δηλαδή που έρχεται αυτόματα να κάμει όλη τη δουλειά για μας. Ο χρήστης χωρίς να γνωρίζει τι κάμνει δεν μπορεί να προχωρήσει παραπέρα με τη δημιουργία οντοτήτων.

Επίσης μέσα από τη μελέτη χρηστών στο δεύτερο μέρος του κεφαλαίου , βλέπουμε ένα εργαλείο να έχει αναπτυχθεί σωστά, αλλά ωστόσο αφήνει περισσότερο από αρκετό χώρο για μελλοντικές βελτιώσεις πάνω στο εργαλείο ώστε να γίνει ακόμα καλύτερο.

Κεφάλαιο 7

ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

7.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα συνεχίσουμε να βλέπουμε το εργαλείο OntoGen σε δράση. Συγκεκριμένα φτιάξαμε τις δικές μας βάσεις από δεδομένα, μέσω ενός λογισμικού εργαλείου το οποίο είναι κατάλληλο γι αυτό το σκοπό, το EndNote X, και ακολούθως τις εφαρμόσαμε μια - μια στο σύστημα μας εξάγοντας τα δικά μας συμπεράσματα.

7.2 Δημιουργία Βάσεων Δεδομένων

Όπως είδαμε και σε προηγούμενα κεφάλαια το εργαλείο OntoGen έχει τη δυνατότητα διαχείρισης μεγάλων συλλογών από κείμενα τα οποία βρίσκονται κάτω από κάποιο κοινό θέμα. Το εργαλείο με τις λειτουργικότητες που διαθέτει πετυχαίνει να τα ταξινομεί σε κατηγορίες και μάλιστα εύκολα αλλά και γρήγορα.

Για τους δικούς μας σκοπούς αξιολόγησης του εργαλείου, αλλά και έρευνας, δημιουργήσαμε τις δικές μας βάσεις δεδομένων από κείμενα στις περιοχές ενδιαφέροντος μας, που είναι και το θέμα που μελετάμε, από βιοιατρική βιβλιογραφία από τη βάση PubMed Central.

Δημιουργήσαμε έξι διαφορετικές βάσεις δεδομένων με τη βοήθεια του εργαλείου EndNote X. Πως δημιουργήσαμε τις βάσεις δεδομένων θα βρείτε τις λεπτομέρειες στο σχετικό παράρτημα στο τέλος της εργασίας (βλέπετε **Παράρτημα Β**).

Οι θεματικές ενότητες με βάση τις οποίες έγιναν η δημιουργία των βάσεων δεδομένων, είναι:

- Autism
- Wireless telemedicine emergency
- Limbic Encephalitis
- Medical imaging and Neural Networks
- Wireless telemedicine emergency, wireless telemedicine ambulance, wireless telemedicine disaster, wireless ambulance, wireless disaster, and wireless emergency
- Biosignals and Neural Networks

7.3 Εφαρμογές

Προχωρούμε έτσι σε αυτή την ενότητα με τις υλοποιήσεις. Όπου κάθε φορά θα τρέχουμε το εργαλείο OntoGen με μια από τις διαφορετικές βάσεις που έχουμε δημιουργήσει και θα σχολιάζουμε και θα αναλύουμε τα αντίστοιχα αποτελέσματα.

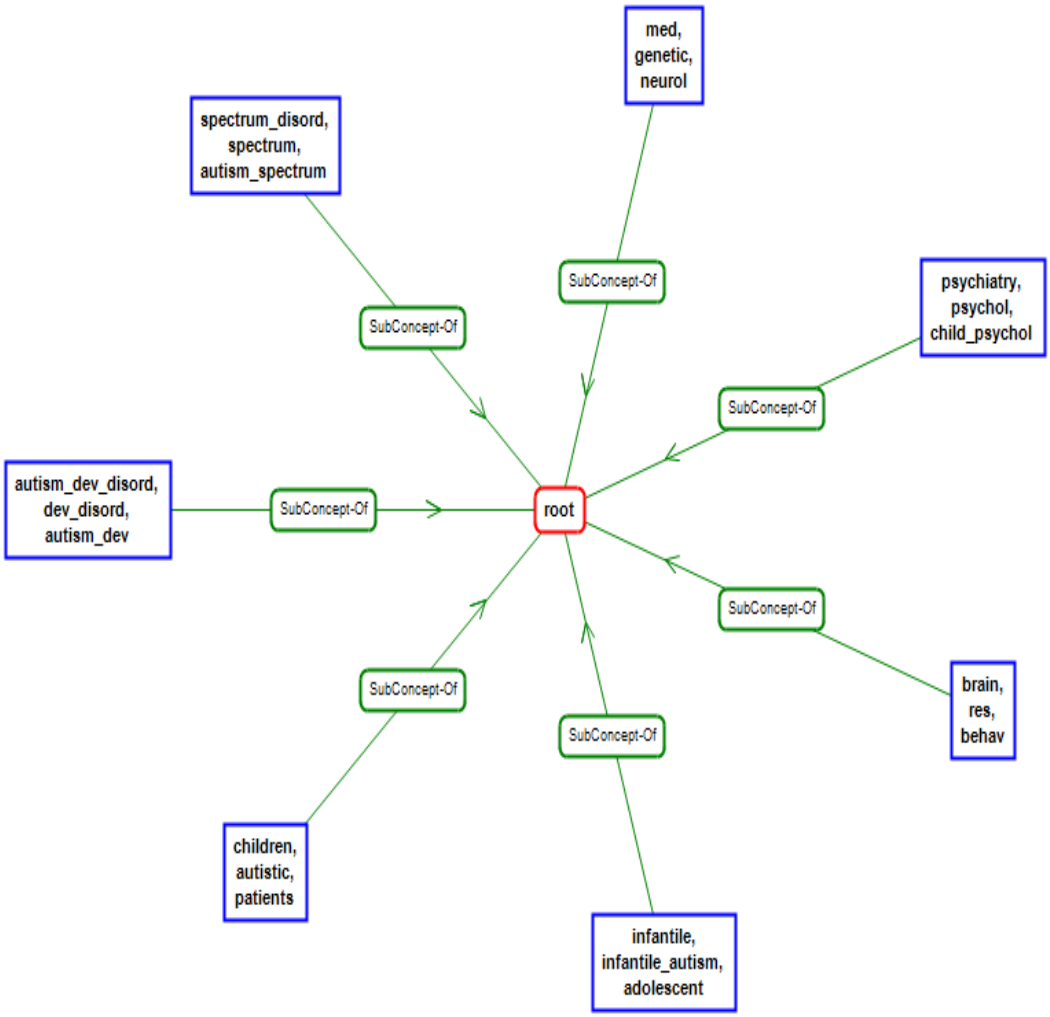
Εφαρμογή 1η

Η βάση δεδομένων που φορτώνεται στο εργαλείο OntoGen είναι η **autism**.txt που αποτελείται από 8491 έγγραφα που ανακτήθηκαν από τη βάση PUBMED.

Στόχος σε κάθε εφαρμογή είναι πρώτα να γίνει μια αναθεώρηση της βιβλιογραφίας και να ανευρεθούν τα θέματα που συναντώνται πιο συχνά μέσα στη βάση που εισάγουμε στο σύστημα. Με αυτό το σκεπτικό παράγεται η οντολογία σε πρώτο

επίπεδο, όπου σε αυτό συναντώνται οι κύριες έννοιες που περιγράφουν τον αυτισμό ή τα φαινόμενα αυτισμού.

Αυτός είναι και ένας τρόπος να έχουμε μια γρήγορη διαδικασία περίληψης και κατανόησης των σύνθετων επιστημονικών άρθρων, όπως φαίνεται και μέσα από το Σχήμα 7.1 που παράγεται από το εργαλείο OntoGen.



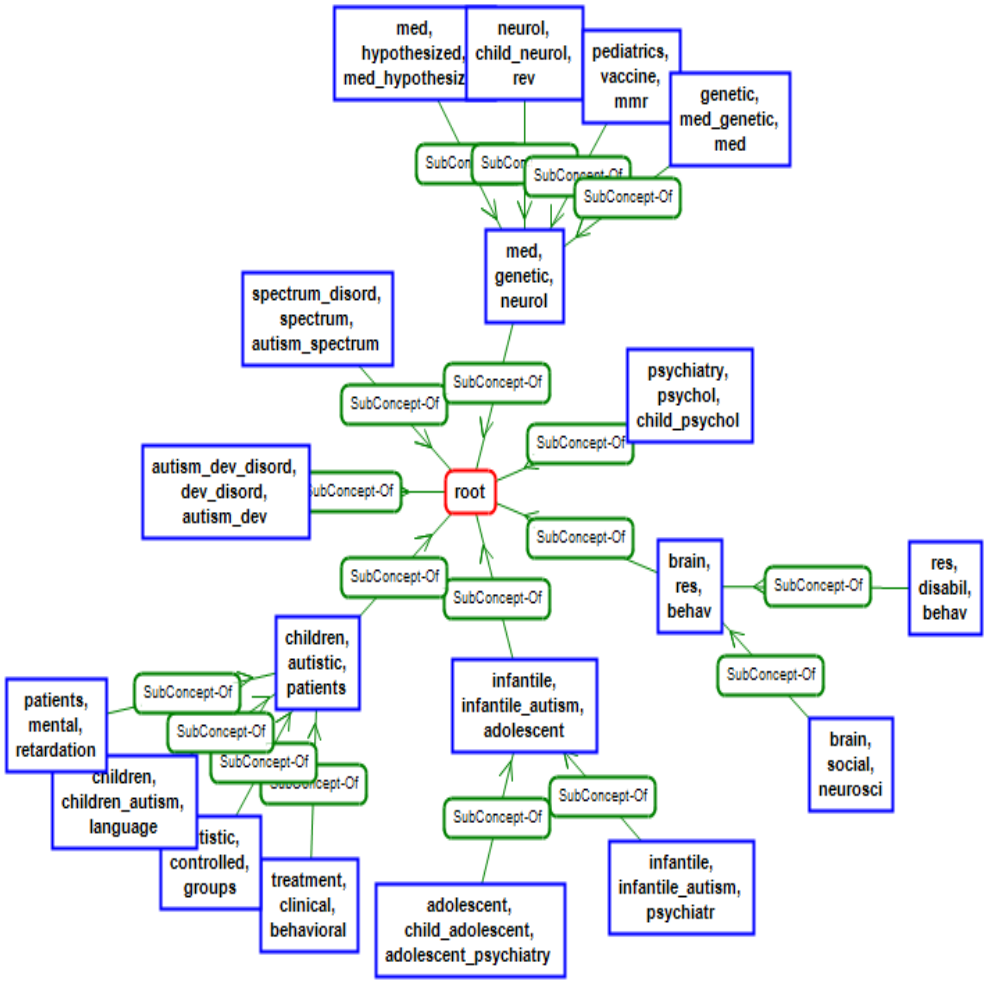
Σχήμα 7.1: Οι έννοιες της οντολογίας αυτισμού με 7 ομάδες, οι οποίες παράγονται από 8491 περιλήψεις κειμένων από τη βάση PubMed Central.

Δηλαδή για να εξηγήσουμε καλύτερα τι εννοούμε με το πιο πάνω αλλά και σε συνδυασμό με το σχήμα, ο χρήστης δημιουργώντας τις πρώτες έννοιες της οντολογίας μπορεί μέσα σε πολύ λίγο χρόνο να πάρει το βασικό νόημα της οντολογίας μας και πιο συγκεκριμένα των κειμένων που την απαρτίζουν.

Για παράδειγμα, αν πάρουμε την έννοια **med, genetic, neurology** πριν καν μελετήσουμε οτιδήποτε μπορούμε αυτόματα να συμπεράνουμε, που αυτός είναι και ένας από τους στόχους της εξόρυξης κειμένων, ενδιαφέρουσες υποθέσεις γύρω από αυτή την έννοια σε σχέση με το θέμα που μελετάμε φυσικά, δηλαδή τον αυτισμό.

Βεβαίως για μια πιο εις βάθος ανάλυση της συγκεκριμένης έννοιας, επιλέγουμε την έννοια και μελετάμε τα έγγραφα (ή στιγμιότυπα) που ανήκουν στην έννοια αυτή.

Προχωρώντας με τη βοήθεια των δύο αλγορίθμων κατηγοριοποίησης (LSI ή k-means), δημιουργούμε την οντολογία και παράγεται ο οντολογικός χάρτης όπως απεικονίζεται στο Σχήμα 7.2.

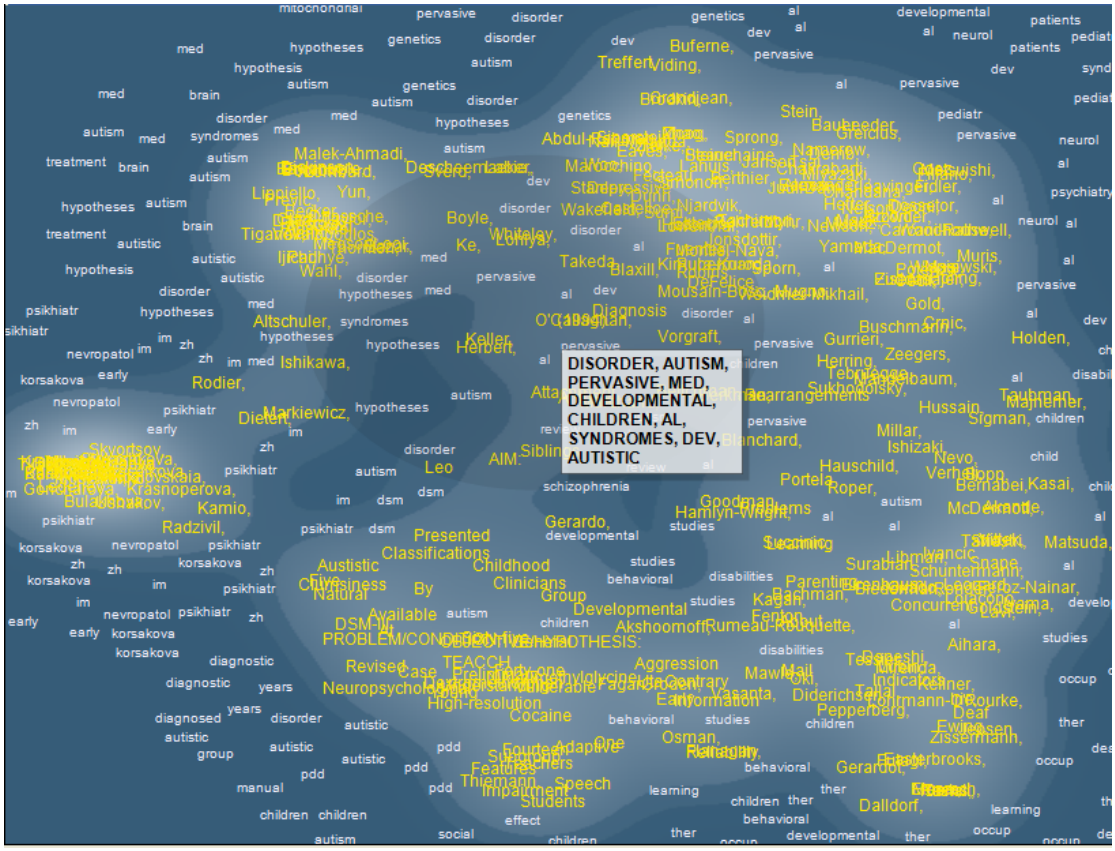


Σχήμα 7.2: Θεματική οντολογία η οποία δημιουργήθηκε από επιστημονικά κείμενα που αφορούν τον αυτισμό.

Η μορφή του Οντολογικού χάρτη μπορεί να έχει διαφορετικές μορφές για την ίδια βάση δεδομένων και αυτό διότι οι αρχικές αποφάσεις αλλά και η τελική μορφή του χάρτη εναπόκεινται όλα στην τελική κρίση του χρήστη. Ο χρήστης είναι αυτός που αποφασίζει εάν μια έννοια θα προστεθεί στην οντολογία, εάν η οντολογία έχει ολοκληρωθεί κ.ο.κ.

Παρ' όλα αυτά η επεξεργασία των δεδομένων εισόδου δεν σταματά εδώ και ο Οντολογικός Χάρτης δεν είναι η μόνη έξοδος του συστήματος από την οποία μπορούμε να εξάγουμε τα συμπεράσματα μας.

Βοηθητική εντοπίζεται να είναι και η οπτική αναπαράσταση των εννοιών (βλέπε Σχήμα 7.3).



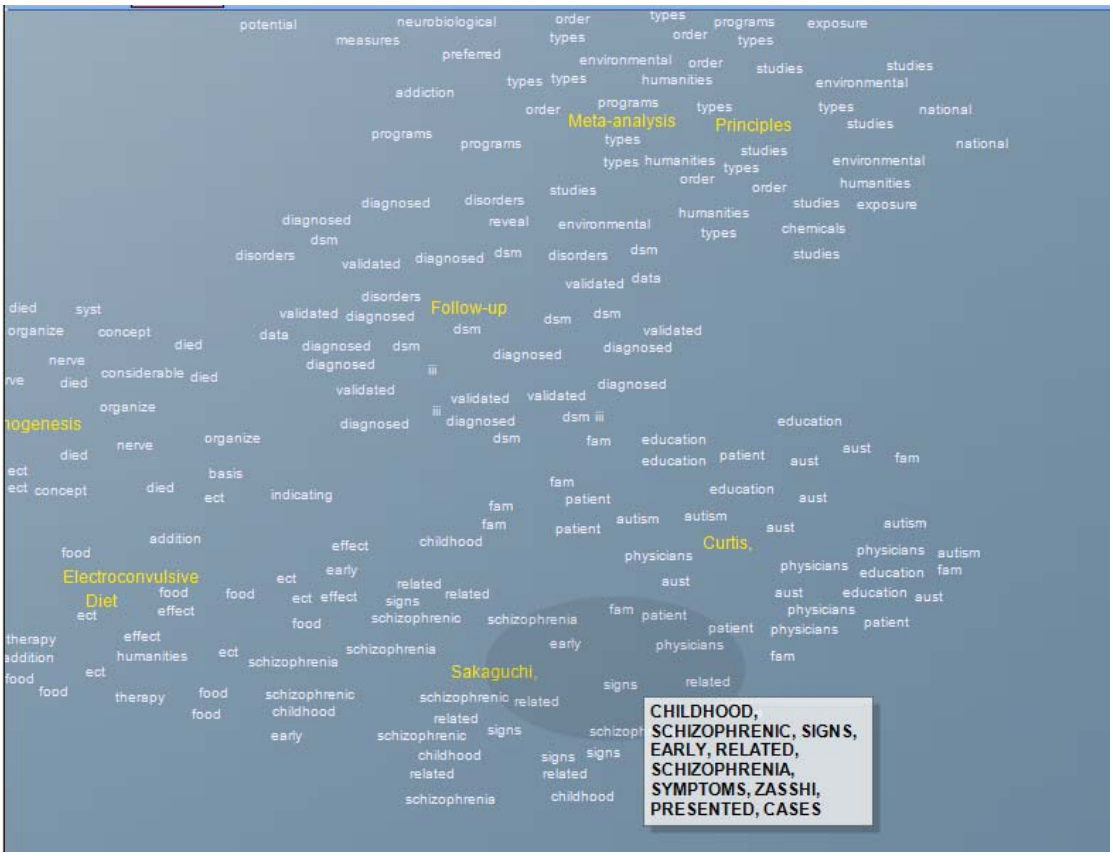
Σχήμα 7.3: Οπτική απεικόνιση έννοιας όπου κάθε στιγμιότυπο ανήκει στην έννοια της οντολογίας developmental, developmental_disorder. Τα στιγμιότυπα όπως φαίνεται από το χάρτη ξεχωρίζουν σε τρεις με τέσσερις ομάδες.

Οι ομάδες στο Σχήμα 7.3 δεν είναι ευδιάκριτες αντιθέτως παρουσιάζουν τομές μεταξύ τους και αυτό διότι οι ομάδες παρουσιάζουν ομοιότητες μεταξύ τους. Αν η ομάδα δεν τέμνεται με μια άλλη ομάδα αυτό υποδηλώνει ότι δεν παρουσιάζουν κανένα κοινό στοιχείο μεταξύ τους.

Με την παροχή αυτή από το σύστημα, της Οπτικής αναπαράστασης κάποιας έννοιας, δίδεται η ευκαιρία στο χρήστη να μελετήσει ακόμα καλύτερα κάποια έννοια. Απομονώνει την έννοια που τον ενδιαφέρει να μελετήσει, επιλέγοντας την από τον Οντολογικό Χάρτη και στη συνέχεια παράγει την Απεικόνιση της έννοιας όπως την είδαμε στο πιο πάνω σχήμα.

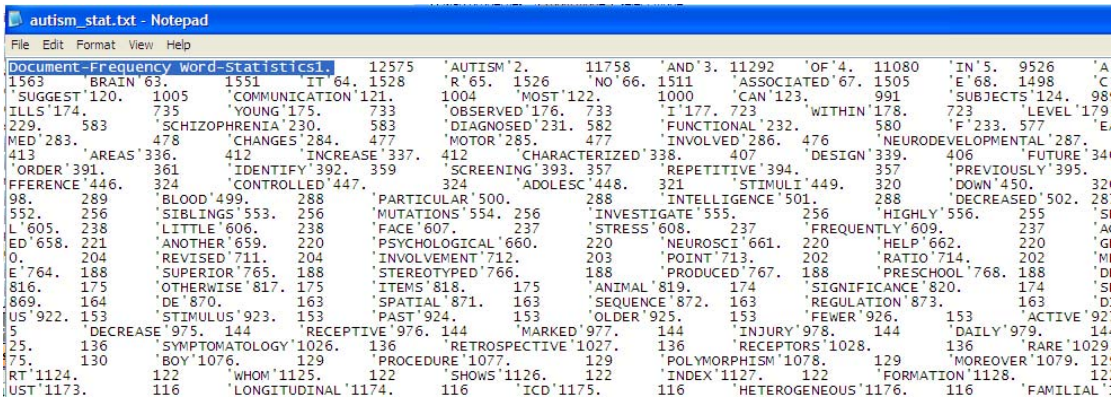
Στο χάρτη αυτό, απεικονίζονται τα στιγμιότυπα της έννοιας σαν σημεία με τέτοιο τρόπο έτσι ώστε κάθε στιγμιότυπο να τοποθετείται κοντά σε παρόμοια στιγμιότυπα και μακριά από λιγότερο όμοια στιγμιότυπα.

Επίσης μπορούμε να επιλέξουμε κάποια περιοχή και να την μεγεθύνουμε για πιο εξονυχιστική μελέτη της έννοιας.



Σχήμα 7.4: Μεγέθυνση μιας περιοχής του χάρτη.

Οι πιο ενδιαφέρουσες υποθέσεις και αποτελέσματα όμως οι οποίες ανακαλύπτονται αυτόματα με τεχνικές εξόρυξης δεδομένων από το σύστημα OntoGen, είναι εκτός από την οντολογία που δημιουργείται από άρθρα από τη βάση PubMed, και το αρχείο *.txt.stat που παράγεται το οποίο περιέχει στατιστικά στοιχεία των όρων όπως εμφανίζονται στα έγγραφα των δεδομένων εισόδου. Δείγμα του αρχείου αυτού βλέπετε στο Σχήμα 7.5. Συγκεκριμένα στο αρχείο αυτό έχουμε κάθε λέξη που εμφανίζεται στα έγγραφα που ανήκουν στη βάση δεδομένων και δίπλα της αναγράφεται ένας αριθμός ο οποίος υποδηλώνει τη συχνότητα εμφάνισης της λέξης αυτής στα δεδομένα.



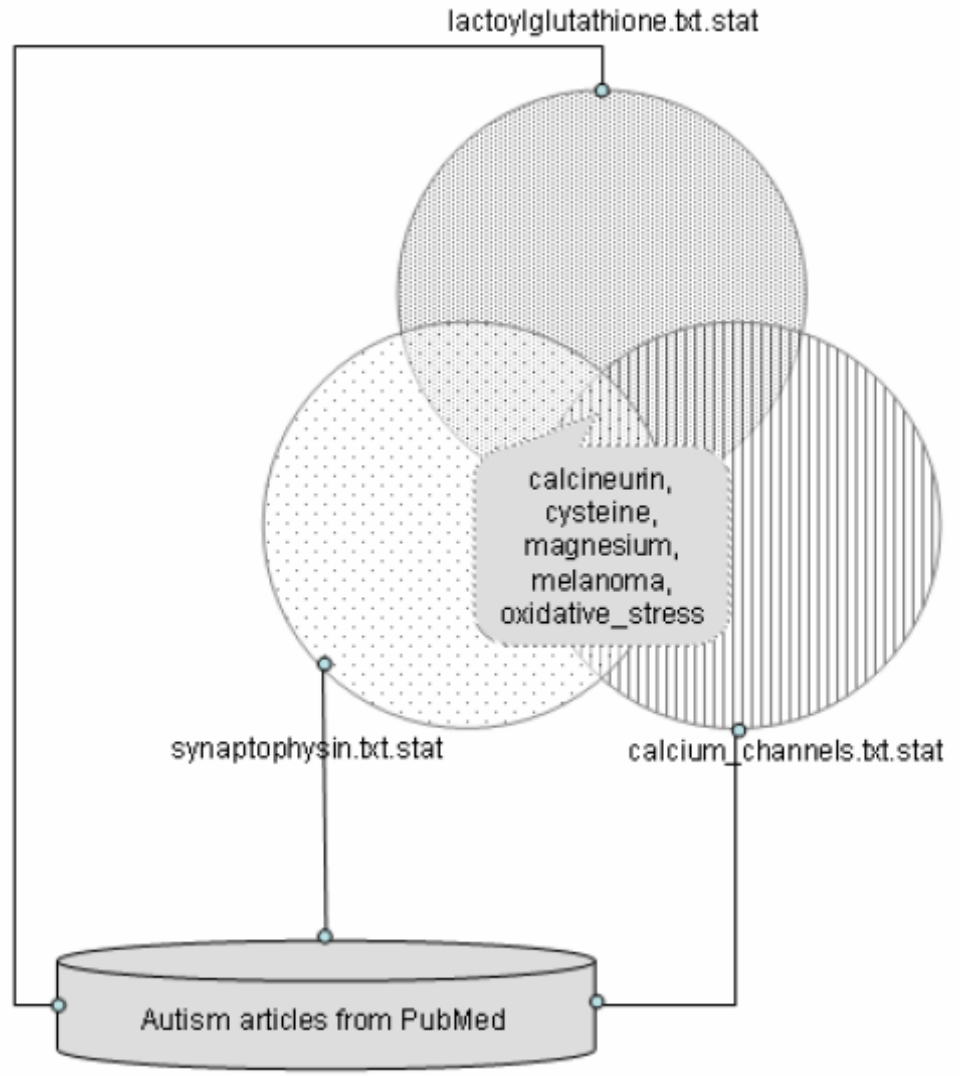
Σχήμα 7.5: Το autism.stat.txt file, το οποίο αναγράφει πόσες φορές παρουσιάζεται ένας όρος στα έγγραφα της βάσης δεδομένων που χρησιμοποιήσαμε για να δημιουργήσουμε την οντολογία (δεδομένα εισόδου).

Το αρχείο αυτό έχει στόχο την ανακάλυψη γνώσης για τον αυτισμό ή για οποιοδήποτε θέμα ερευνάμε, η οποία είναι ατεκμηρίωτη. Έτσι ξεκινώντας την έρευνα μας πάνω σε σπάνιους συνδέσεις μεταξύ των δεδομένων παρά σε πιο συχνούς εμφανιζόμενους όρους, θα έχουμε μεγαλύτερες πιθανότητες να ανακαλύψουμε κάποιες υπονοούμενες γνώσεις οι οποίες είναι ακόμα άγνωστες και μπορεί να είναι χρήσιμες για τους ερευνητές.

Για να γίνει η πιο πάνω έρευνα όμως πρέπει να τονιστεί ότι απαιτείται από το χρήστη που συνήθως είναι κάποιος ερευνητής, κάποια υπόβαθρο στο συγκεκριμένο θέμα που ερευνά. Για παράδειγμα ένας γιατρός μπορεί άνετα να χειριστεί το εργαλείο για να

μελετήσει ιατρικά θέματα τα οποία δεν έχουν κατανοηθεί πλήρως από διάφορες απόψεις.

Παρακάτω παρουσιάζουμε τα αποτελέσματα όπως εξήχθησαν από το *.txt.stat file.



Σχήμα 7.6 : Τα αποτελέσματα από τη εξόρυξη βιβλιογραφίας μέσα στο πεδίο αυτισμού.

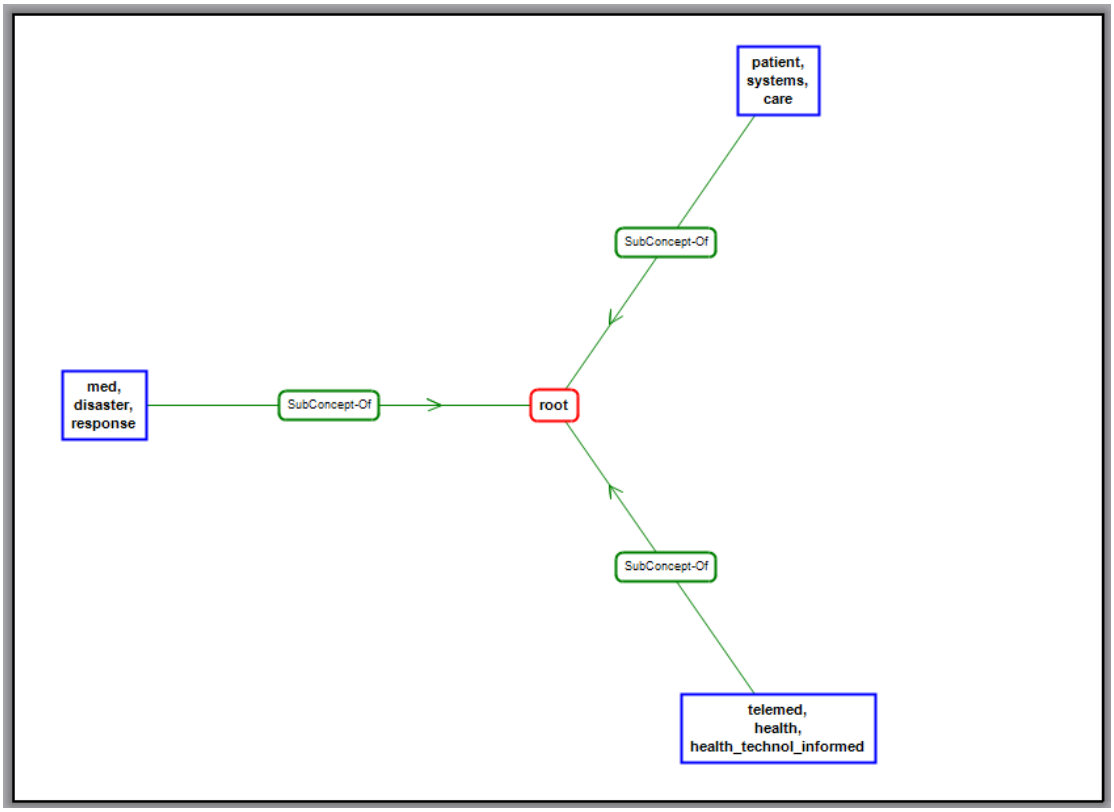
Από τα *.txt.stat files ανακτώνται οι λέξεις οι οποίες είναι κοινές σε όλα τα αρχεία. Ένας από αυτούς τους όρους ο οποίος απεικονίζεται και στο Σχήμα 7.6, ο οποίος θα μπορούσε να ήταν ενδιαφέρον για περαιτέρω έρευνα του αυτισμού, είναι ο calcineurin. Calcineurin είναι calcium και calcium-dependent, το οποίο εντοπίζεται κυρίως στους ιστούς των θηλαστικών, με τα μεγαλύτερα ποσοστά να βρίσκονται στον εγκέφαλο. Η εξόρυξη βιβλιογραφίας έδειξε ότι θα μπορούσε να σχετίζεται με

αυτιστικές διαταραχές, ωστόσο προς το παρόν δεν υπάρχει κάποιο στοιχείο στο διαδίκτυο που να υποδεικνύει το ρόλο του calcineurin στον αυτισμό.

Εφαρμογή 2η

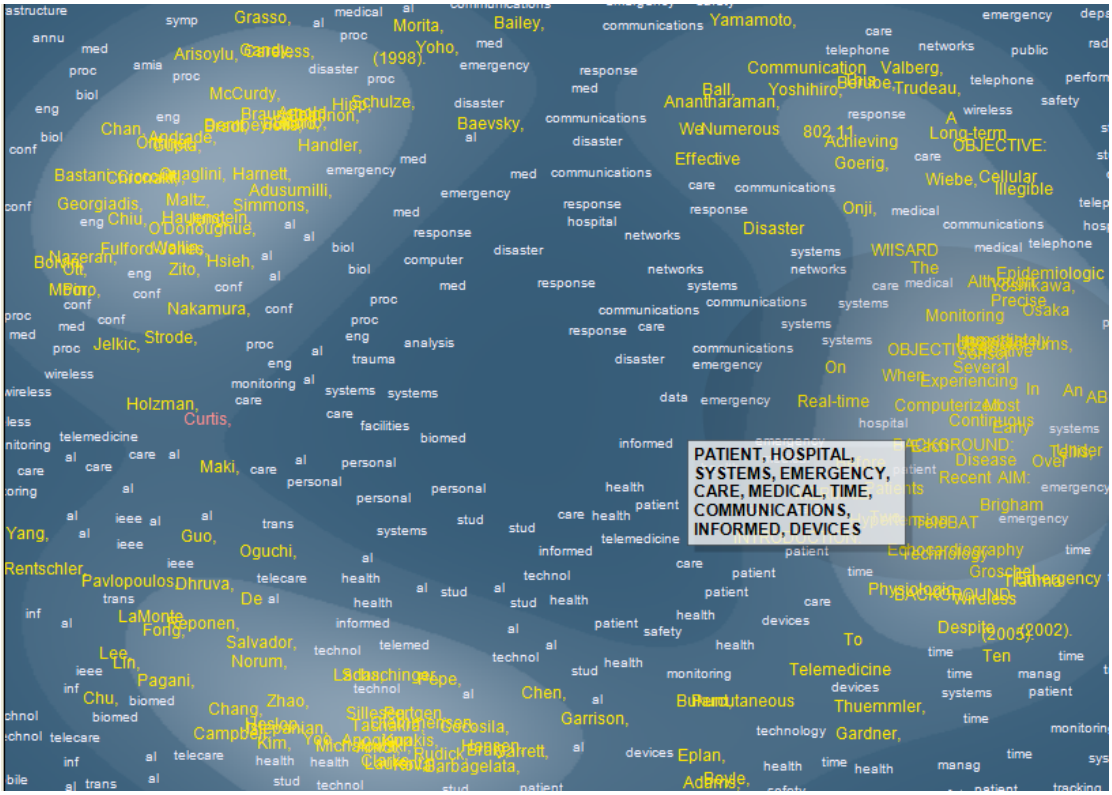
Η βάση δεδομένων που φορτώνεται στο εργαλείο OntoGen είναι η **wireless_emergency**.txt που αποτελείται από 145 έγγραφα που ανακτήθηκαν από τη βάση PUBMED. Στο Σχήμα 7.7 απεικονίζεται ο Οντολογικός Χάρτης που παράγεται μόλις φορτωθεί η βάση δεδομένων. Έτσι έχουμε τις κύριες έννοιες σχετικά με το θέμα που μελετάμε "wireless emergency" όπως προκύπτουν από το πρώτο επίπεδο του οντολογικού μοντέλου, οι οποίες είναι:

- patient, systems, care
- med, disaster, response
- telemedicine, health, health_technology_informed



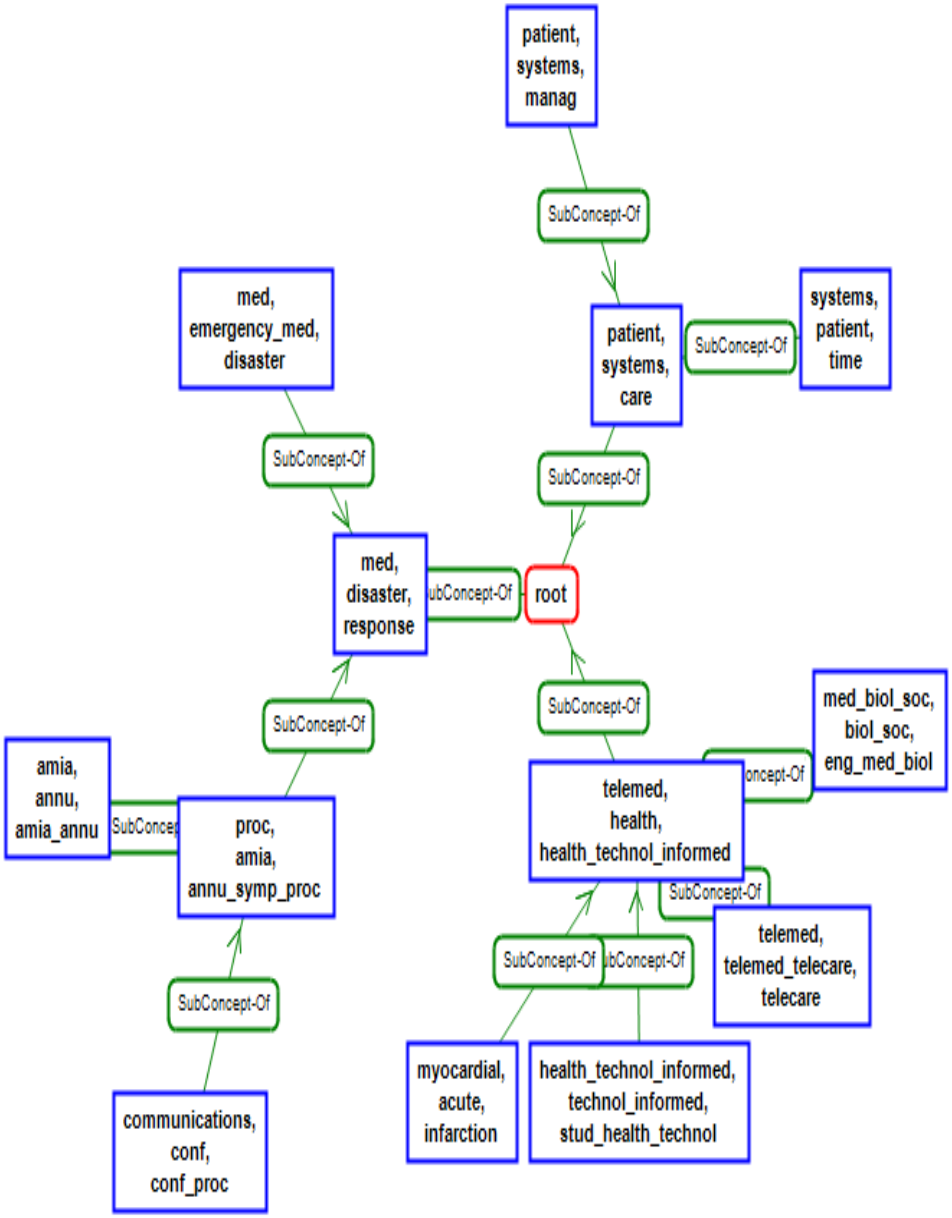
Σχήμα 7.7: Η Θεματική Οντολογία που κατασκευάστηκε με βάση τις περιγραφές κειμένων που είχαμε στη βάση δεδομένων σχετικά με το θέμα "wireless emergency".

Τις ίδιες ομάδες θα διακρίνουμε αν επιλέξουμε τη root έννοια και δούμε την απεικόνιση της, όπως φαίνεται στο Σχήμα 7.8. Το σύστημα διαχωρίζει το root concept σε τρείς αυτόνομες ομάδες, άρα αυτό είναι και βοηθητικό για το χρήστη έτσι ώστε να αποφασίσει στη συνέχεια σε πόσες ομάδες θα κατηγοριοποιήσει τα δεδομένα του, διότι αυτό είναι πάντοτε επιλογή του χρήστη. Για αυτό το λόγο και εγώ επέλεξα τρείς ομάδες και τον αλγόριθμο k-means για να έχουμε τα αποτελέσματα του σχήματος 7.7.



Σχήμα 7.8: Οπτική απεικόνιση του root concept. Διακρίνονται τρείς ξεχωριστές ομάδες.

Με την ίδια τεχνική προχώρησα παρακάτω για τη δημιουργία και ολοκλήρωση της οντολογίας. Έτσι προκύπτει το Σχήμα 7.9.



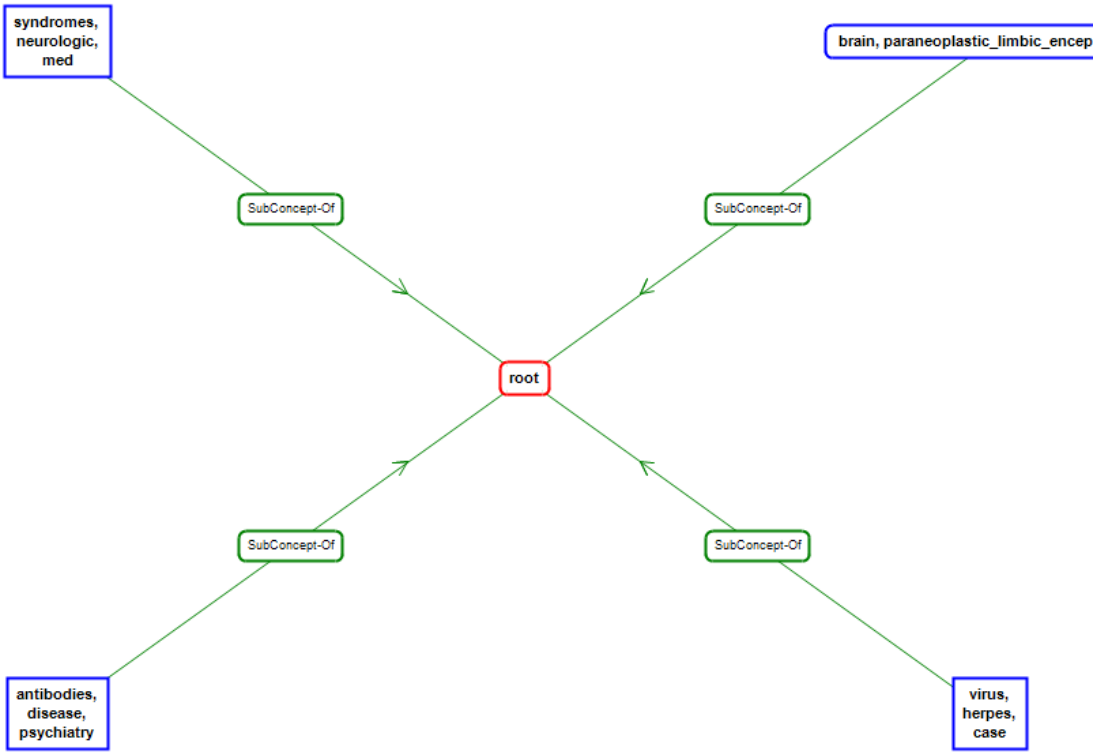
Σχήμα 7.9: Θεματική Οντολογία η οποία δημιουργήθηκε από τις περιγραφές σχετικές με wireless emergency που πάρθηκαν από τη βάση PubMed.

Αν ο ερευνητής επιθυμεί για μια πιο εις βάθος μελέτη για το αντικείμενο που μελετά μπορεί να κάμει χρήση του αρχείου wireless_emergency_stat.txt, με τον τρόπο που αναφέραμε στην πρώτη εφαρμογή.

Εφαρμογή 3η

Η βάση δεδομένων που φορτώνεται στο εργαλείο OntoGen είναι η **limbic_encephalitis.txt** που αποτελείται από 498 έγγραφα που ανακτήθηκαν από τη βάση PUBMED. Στο Σχήμα 7.10 απεικονίζεται ο Οντολογικός Χάρτης που παράγεται μόλις φορτωθεί η βάση δεδομένων. Έτσι έχουμε τις κύριες έννοιες σχετικά με το θέμα που μελετάμε "limbic encephalitis" όπως προκύπτουν από το πρώτο επίπεδο του οντολογικού μοντέλου, οι οποίες είναι:

- brain, paraneoplastic_limbic_encephalitis
- syndromes, neurologic, medicine
- antibodies, disease, psychiatry
- virus, herpes, case



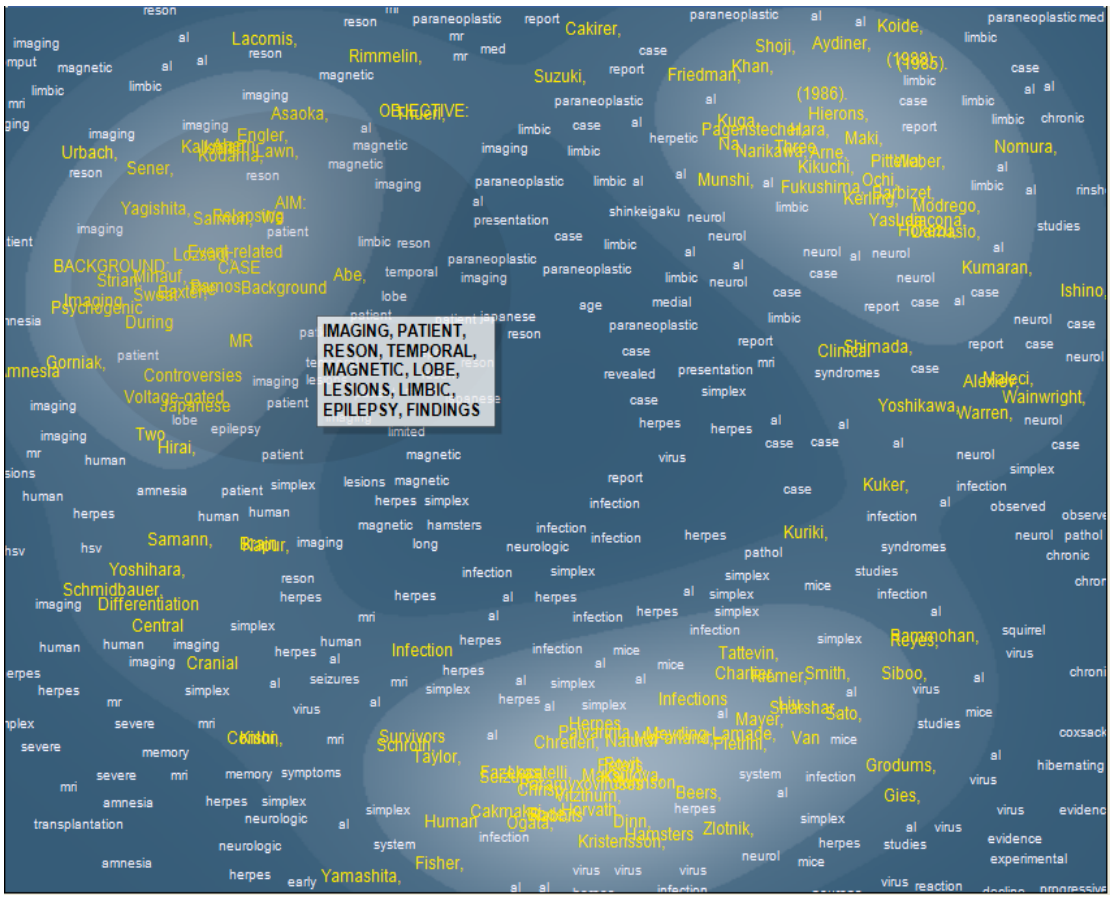
Σχήμα 7.10: Το πρώτο στρώμα της οντολογίας όπως δημιουργείται με επιλογή τεσσάρων παραμέτρων και του αλγορίθμου k-means. Συνεπώς προκύπτουν 4 κύριες έννοιες αρχικά.

Ένας ειδικός που ειδικεύεται στο θέμα αυτό θα μπορούσε πολύ εύκολα να δικαιολογήσει το διαχωρισμό και να πιστοποιήσει την ορθότητα του. Από την άλλη πάλι για ένα πρωτόπειρο που θα ήθελε να ασχοληθεί εξ'αρχής με το θέμα για οποιοδήποτε λόγο έρευνας, θα ήταν μια καλή αρχή.

Όντως αν δοκιμάσουμε να διαβάσουμε εν συντομία το σχήμα 7.10, η limbic encephalitis με απλά λόγια είναι μια ασθένεια που εστιάζεται στον εγκέφαλο και μπορεί να την βρούμε με τη μορφή Paraneoplastic limbic encephalitis ($1^{η}$ έννοια), μπορεί να προκληθεί από κάποιο ιό που προσβάλει τον οργανισμό ($2^{η}$ έννοια), ή πάλι μπορεί να προκληθεί από κάποια βλάβη στον τρόπο δράσης των αντισωμάτων στον οργανισμό – υπολειτουργία ($3^{η}$ έννοια) και τέλος κάποια σύνδρομα, παρενέργειες προκαλούνται από την βλάβη αυτή στον εγκέφαλο. Αυτές συνήθως είναι νευρολογικές ασθένειες. Οι πληροφορίες αυτές προκύπτουν από το σχήμα αλλά κι αν μια έννοια δεν είναι τελείως ξεκάθαρη μπορούμε να συμβουλευτούμε το περιεχόμενο των εγγράφων που ανήκουν σε αυτή.

Στο Σχήμα 7.11 βλέπουμε την οπτική απεικόνιση της έννοιας **virus, herpes, case.** Προκύπτουν τρείς ευδιάκριτες ομάδες, έτσι αποφάσισα για τη συγκεκριμένη έννοια να χρησιμοποιήσω αριθμό εισηγήσεων 3 και τον αλγόριθμο k-means για να προσθέσω τις υπό – έννοιες στην έννοια αυτή.

Κάθε φορά ακολουθούσα αυτή την τακτική για καλύτερα αποτελέσματα. Έτσι τελικά προέκυψε το Σχήμα 7.12.



Σχήμα 7.11: Οπτική απεικόνιση της έννοιας **virus, herpes, case.**

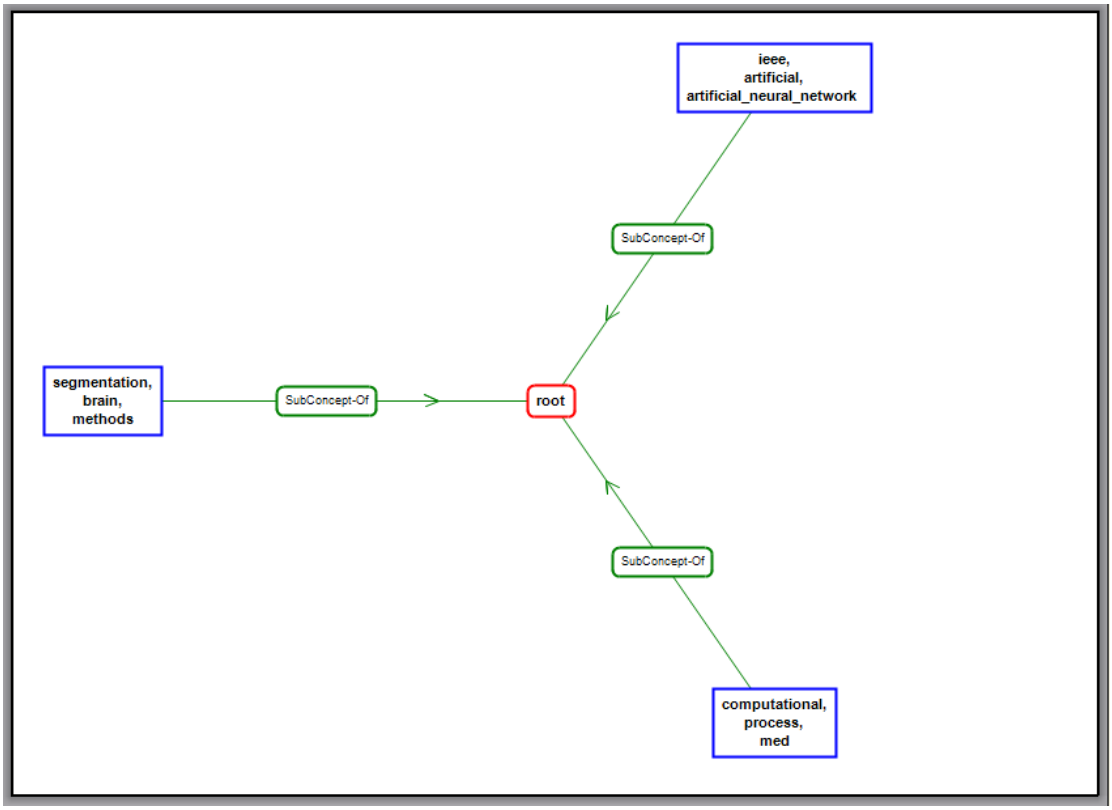
Σχήμα 7.12: Θεματική οντολογία “Limbic Encephalitis”.

Η τελική μορφή της οντολογίας αποτελείται από δύο επίπεδα. Το πρώτο επίπεδο με τις κύριες έννοιες που προαναφέραμε και το δεύτερο επίπεδο με τις υπό – έννοιες για κάθε έννοια όπως αναπαριστώνται στο πιο πάνω σχήμα.

Εφαρμογή 4η

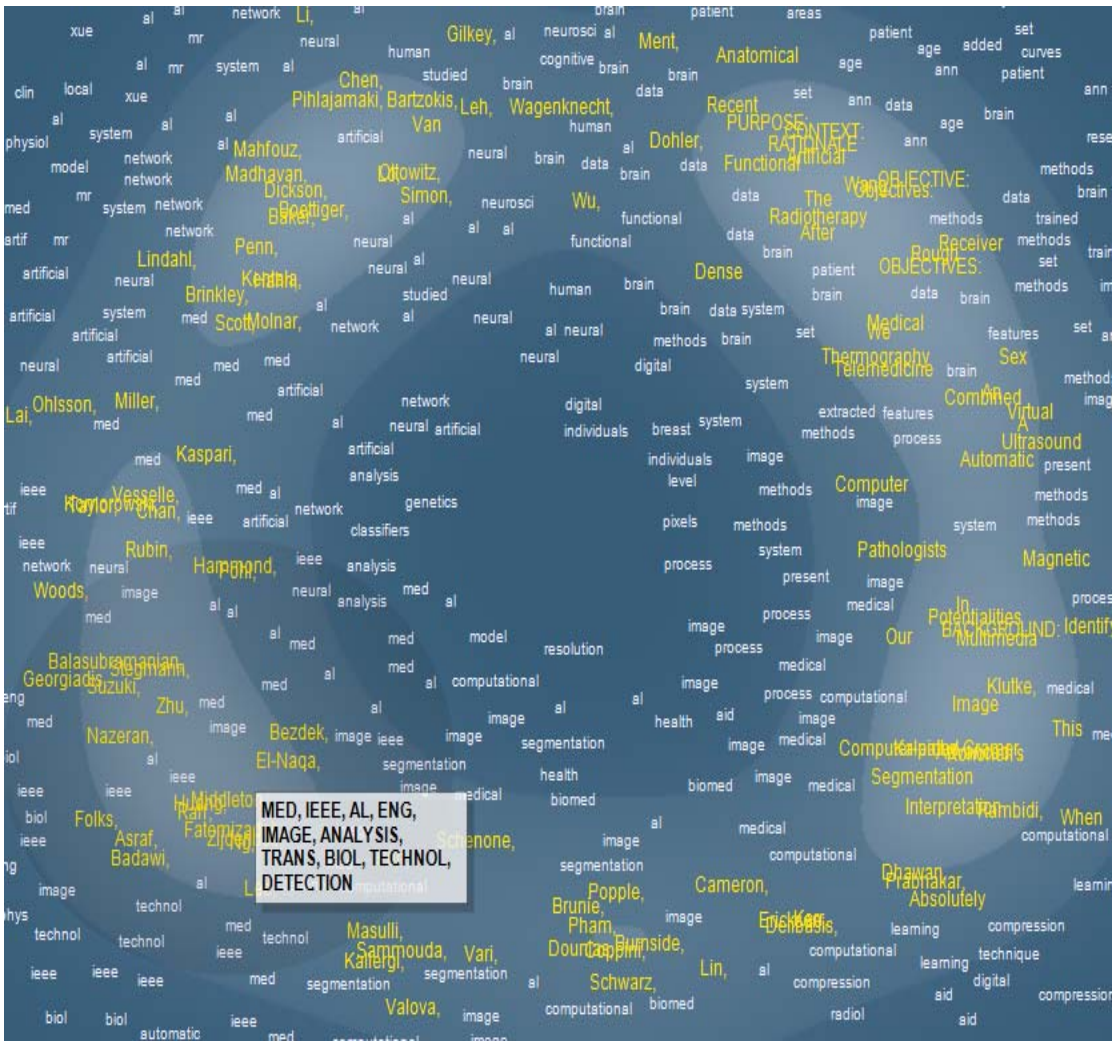
Η βάση δεδομένων που φορτώνεται στο εργαλείο OntoGen είναι η **medical_images_and_neural_networks**.txt που αποτελείται από 89 έγγραφα που ανακτήθηκαν από τη βάση PUBMED. Στο Σχήμα 7.13 απεικονίζεται ο Οντολογικός Χάρτης που παράγεται μόλις φορτωθεί η βάση δεδομένων. Έτσι έχουμε τις κύριες έννοιες σχετικά με το θέμα που μελετάμε "medical images and neural networks" όπως προκύπτουν από το πρώτο επίπεδο του οντολογικού μοντέλου, οι οποίες είναι:

- artificial neural network
- segmentation, brain, methods
- computational, process, med



Σχήμα 7.13: Το πρώτο επίπεδο της Θεματικής οντολογίας με τις 3 βασικές έννοιες όπως προκύπτουν από τον αλγόριθμο LSI.

Η βασική έννοια "root" κατηγοριοποιήθηκε σε τρείς υπό – κατηγορίες αφού πρώτα παράξαμε τη Σχήμα 7.14, δηλαδή την οπτική απεικόνιση της βασικής έννοιας. Με αυτό τον τρόπο, όπως και στις προηγούμενες εφαρμογές, το σύστημα ομαδοποιεί τα στιγμιότυπα με βάση τις ομοιότητες που παρουσιάζουν μεταξύ τους, Συνεπώς με τη βοήθεια του σχήματος 7.14, όπου διακρίνονται εύκολα τρείς ομάδες, αποφασίσαμε να "σπάσουμε" την βασική έννοια σε τρία μέρη.

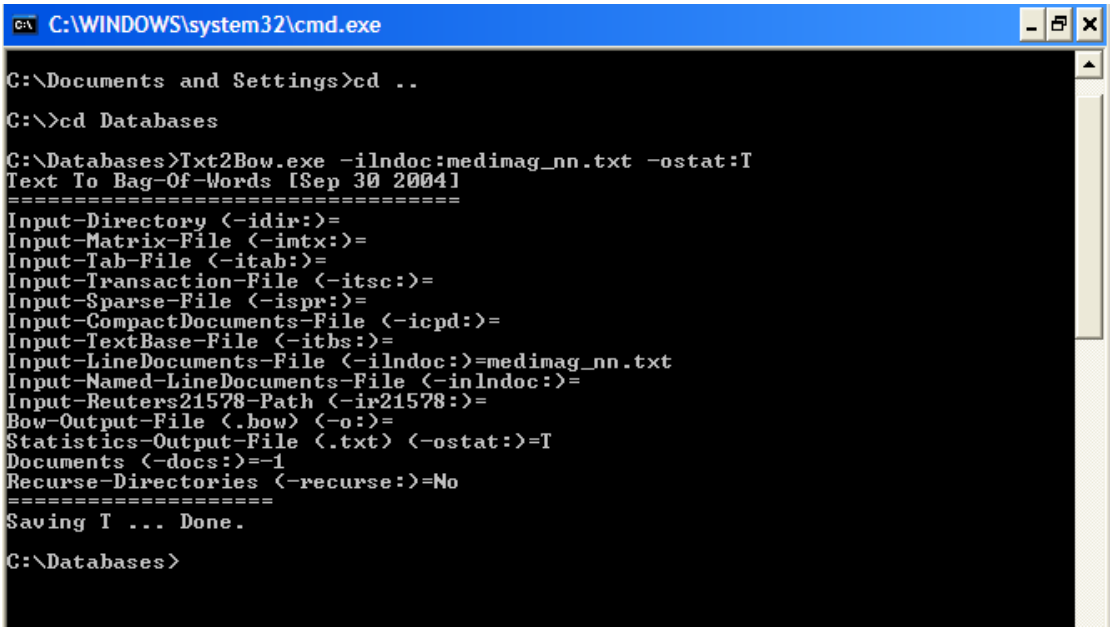


Σχήμα 7.14: Οπτική απεικόνιση της βασικής έννοιας.

Σχήμα 7.15: Ο οντολογικός χάρτης όπως προκύπτει από τη βάση δεδομένων "**medical_images_and_neural_networks**.txt".

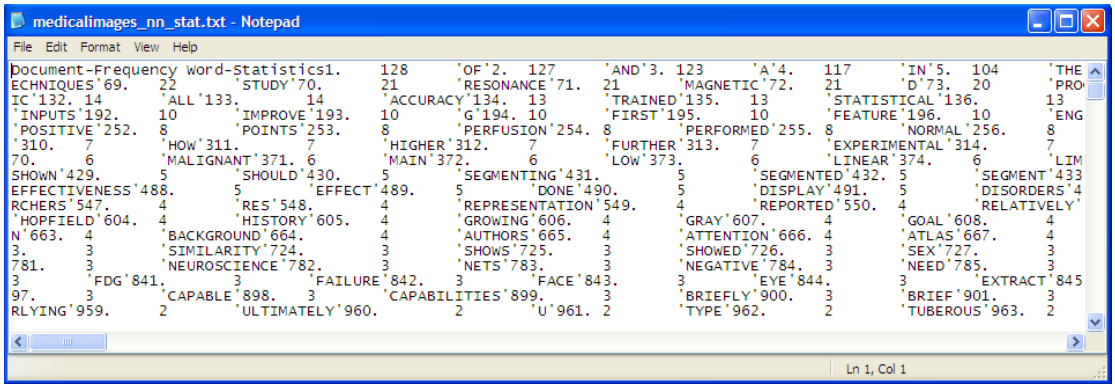
Το Σχήμα 7.15 είναι η τελική μορφή της οντολογίας. Ο χρήστης μπορεί να φτάσει το χάρτη σε οποιαδήποτε μορφή επιθυμεί ανάλογα με το πόσο πολύ θέλει να προχωρήσει στην έρευνα του. Για παράδειγμα η μορφή της οντολογίας του σχήματος 7.15 είναι αρκετά πολύπλοκη και προορίζεται για μια πιο εις βάθος μελέτη από τον ερευνητή. Η οντολογία θα μπορούσε να αποτελείτο ίσως και από ένα ακόμα επίπεδο. Αυτό όμως εξαρτάται και σε κάποιο βαθμό από το μέγεθος της βάσης δεδομένων μας.

Παραθέτουμε ακόμη ένα παράδειγμα που υποδεικνύει τον τρόπο παραγωγής του *.stat.txt file όπως φαίνεται στο Σχήμα 7.16.



Σχήμα 7.16: Στη γραμμή εκτέλεσης βάζουμε σαν είσοδο το αρχείο εισόδου και με την εντολή Txt2Bow.exe και την παράμετρο –ostat το αρχείο αυτό μετατρέπεται σε αρχείο με στατιστικά δεδομένα.

Στο Σχήμα 7.17 απεικονίζεται το αρχείο που παράγει η πιο πάνω εκτέλεση.



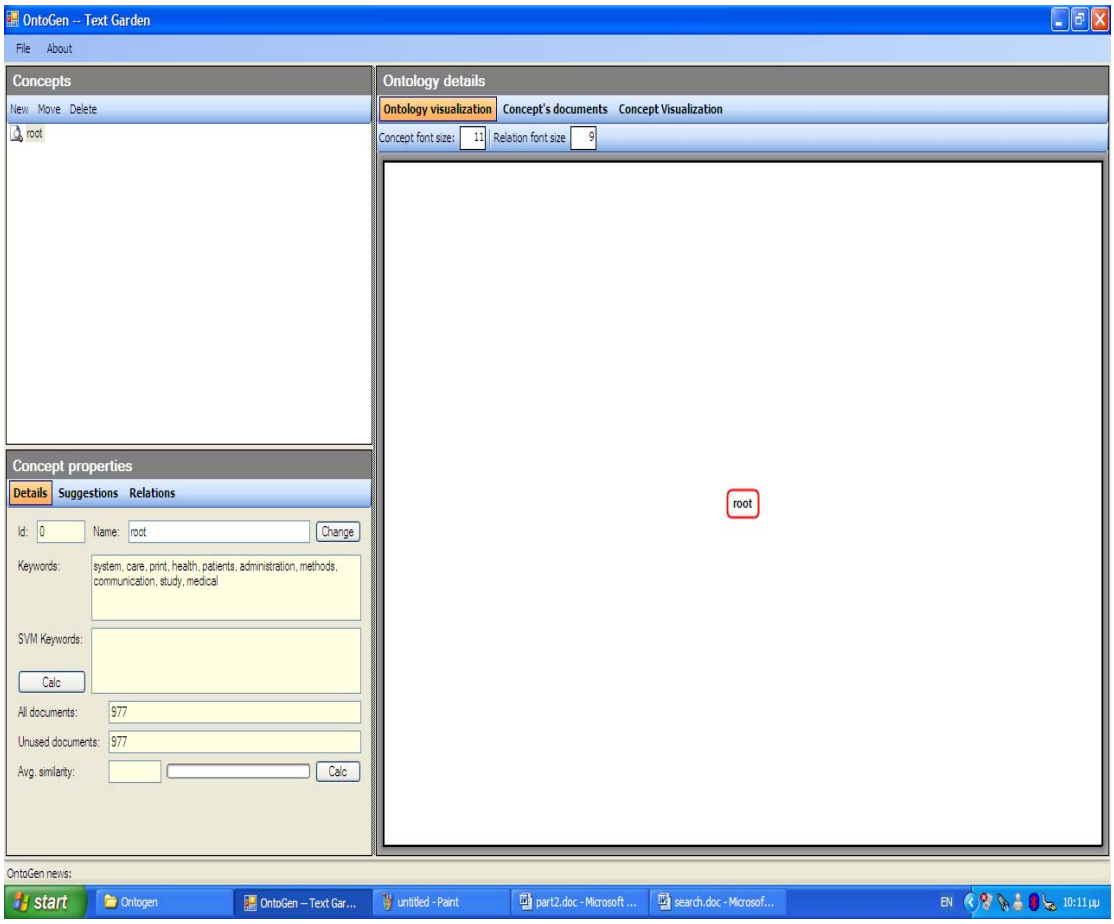
Σχήμα 7.17: Το αρχείο "medicalimages_nn_stat.txt".

Εφαρμογή 5η

Η βάση δεδομένων που δημιουργήσαμε σε αυτή την περίπτωση αποτελείται από 977 έγγραφα και ανακτήθηκαν από τη βάση PUBMED, κάνοντας αναζήτηση για τους πιο κάτω συγκεκριμένους όρους:

- Wireless telemedicine emergency, wireless telemedicine ambulance, wireless telemedicine disaster, wireless ambulance, wireless disaster, and wireless emergency.

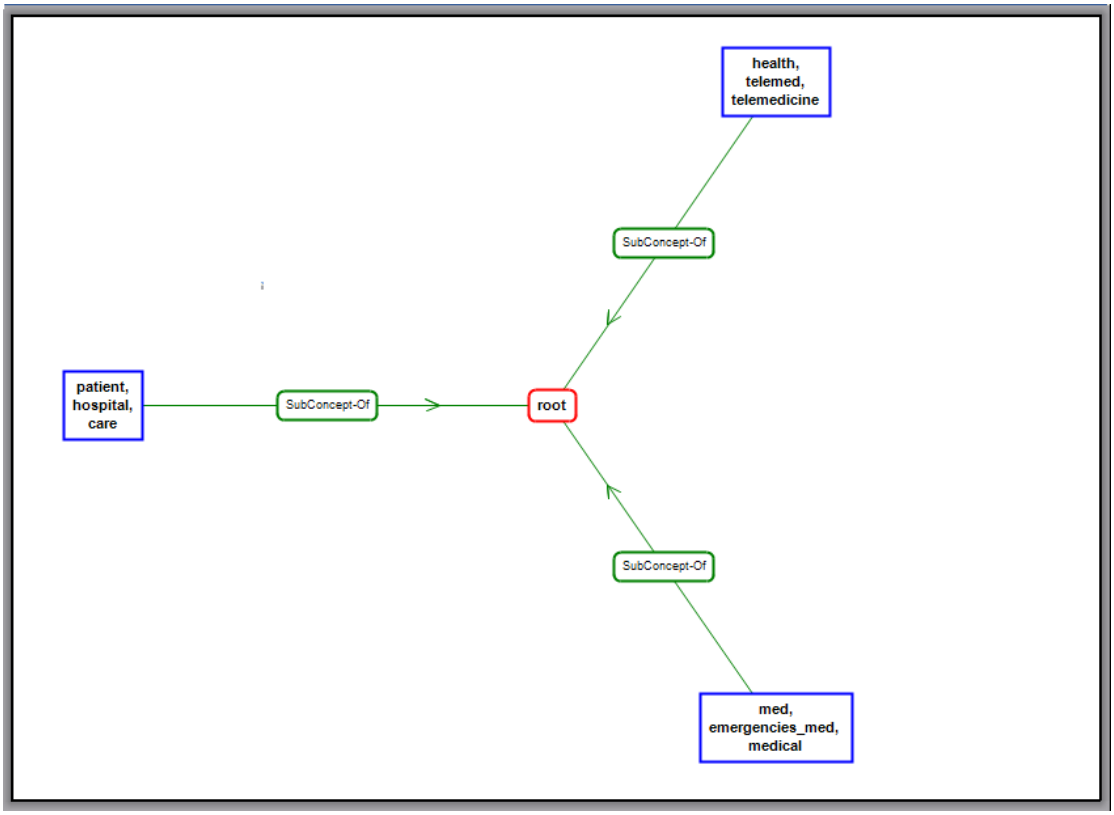
Η διαδικασία που ακολουθείται για τη δημιουργία μιας Οντολογίας από το σύστημα OntoGen, είναι πάντοτε η ίδια. Έτσι και για αυτή τη φορά ως πρώτο βήμα φορτώνουμε τα δεδομένα μας (τη βάση δεδομένων που έχουμε δημιουργήσει), και αυτόματα στην οθόνη του συστήματος θα εμφανιστεί η βασική έννοια και οι κύριες λέξεις κλειδιά που την περιγράφουν (βλέπε Σχήμα 7.18).



Σχήμα 7.18: Στο δεξί παράθυρο απεικονίζεται η βασική έννοια της οντολογίας και στο κάτω αριστερά παράθυρο, φαίνονται οι κύριες λέξεις κλειδιά που περιγράφουν την επιλεγμένη έννοια (τη βασική έννοια).

Επόμενο βήμα για την δημιουργία την οντολογίας είναι η εισήγηση εννοιών για την πρόσθεση εννοιών και στη συνέχεια υπό – εννοιών στην οντολογία. Ο χρήστης μπορεί να επιλέξει με ποια μέθοδο μπορεί να γίνει η εισήγηση. Είτε με ένα από τους δύο αλγόριθμους κατηγοριοποίησης (LSI ή k-means) ή με εισαγωγή ενός query. Έτσι βήμα προς βήμα, με τη καθοδήγηση του συστήματος αλλά και την κρίση του χρήστη η οντολογία ολοκληρώνεται.

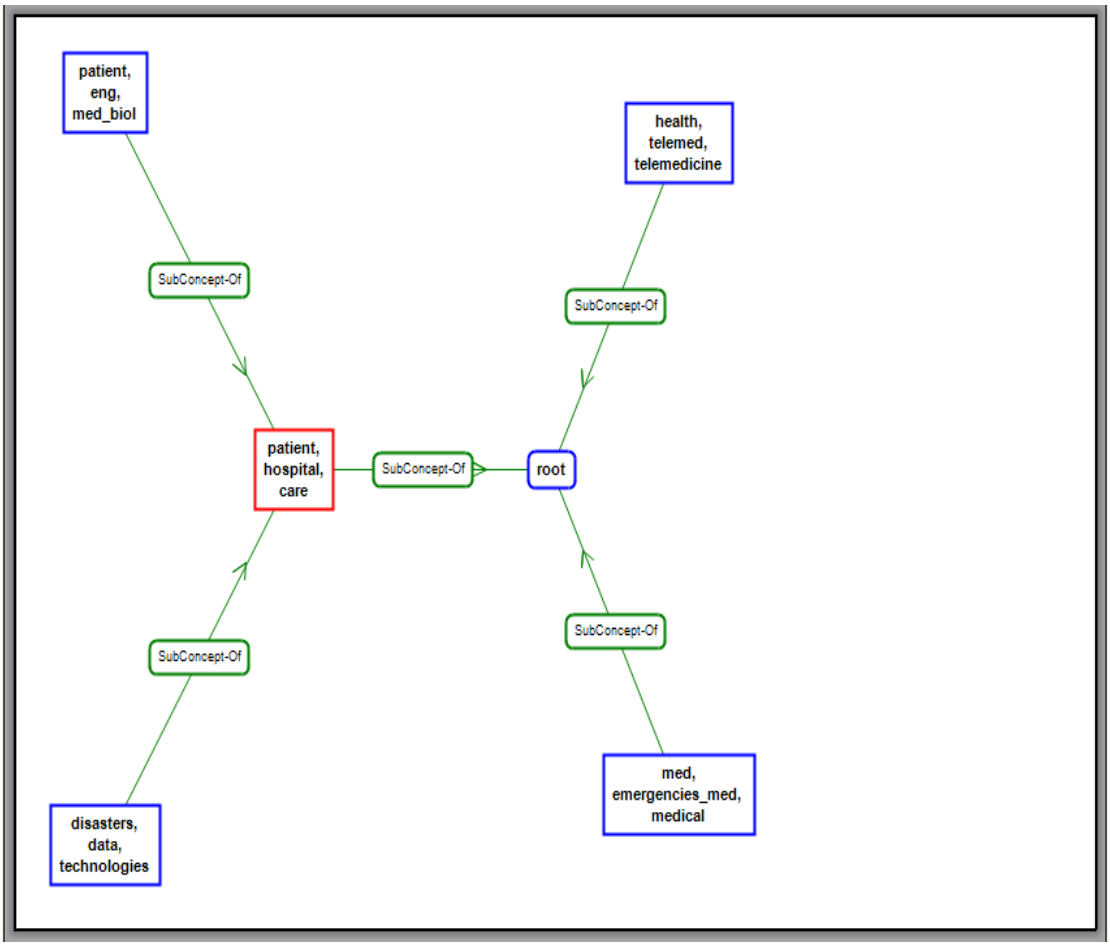
Στο Σχήμα 7.19, πιο συγκεκριμένα ο χρήστης επιλέγει τον αλγόριθμο LSI και αριθμό εισηγήσεων 3, γι' αυτό το λόγο προκύπτουν και τρείς έννοιες στην οπτική απεικόνιση της οντολογίας.



Σχήμα 7.19: Η οπτική απεικόνιση του πρώτου στρώματος από έννοιες στην οντολογία.

Στη συνέχεια, ο χρήστης μπορεί να επιλέξει μια έννοια από την οπτική απεικόνιση της οντολογίας ή ακόμα και από την ιεραρχία της οντολογίας, τον άλλο τρόπο αναπαράστασης της οντολογίας, και να προσθέσει με τον τρόπο που είδαμε στο πιο πάνω βήμα υπό – έννοιες για αυτή την έννοια.

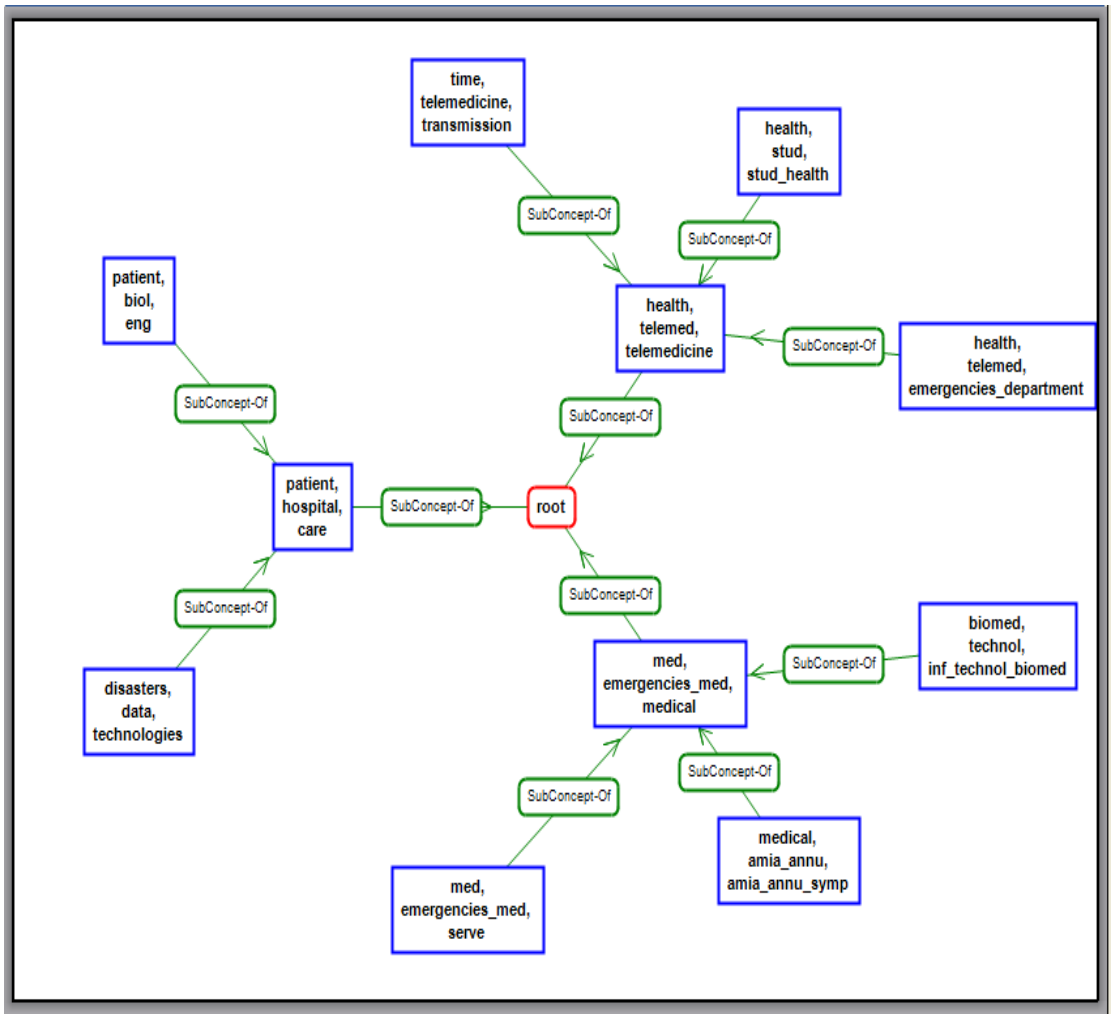
Για παράδειγμα, έστω επιλέγουμε την έννοια “patient, hospital, care”, την οποία το σύστημα χρωματίζει με κόκκινο χρώμα. Προσθέτοντας υπό – έννοιες για την επιλεγμένη έννοια η μορφή της οντολογίας παίρνει τη μορφή του σχήματος 7.20.



Σχήμα 7.20: Η επιλεγμένη έννοια είναι με κόκκινο χρώμα και στο σχήμα φαίνονται οι υπό – έννοιες τις οποίες προσθέσαμε γι αυτή την έννοια.

Ο χρήστης μπορεί να εκτείνει το μέγεθος της οντολογίας του αναλόγως με το μέγεθος της βάσης δεδομένων που επεξεργάζεται αλλά και αναλόγως με το πόσο ενδιαφέρεται να προχωρήσει την έρευνα του στο συγκεκριμένο θέμα.

Έτσι ο χρήστης μπορεί να δημιουργήσει την οντολογία του Σχήματος 7.21 με δύο στρώματα εννοιών.

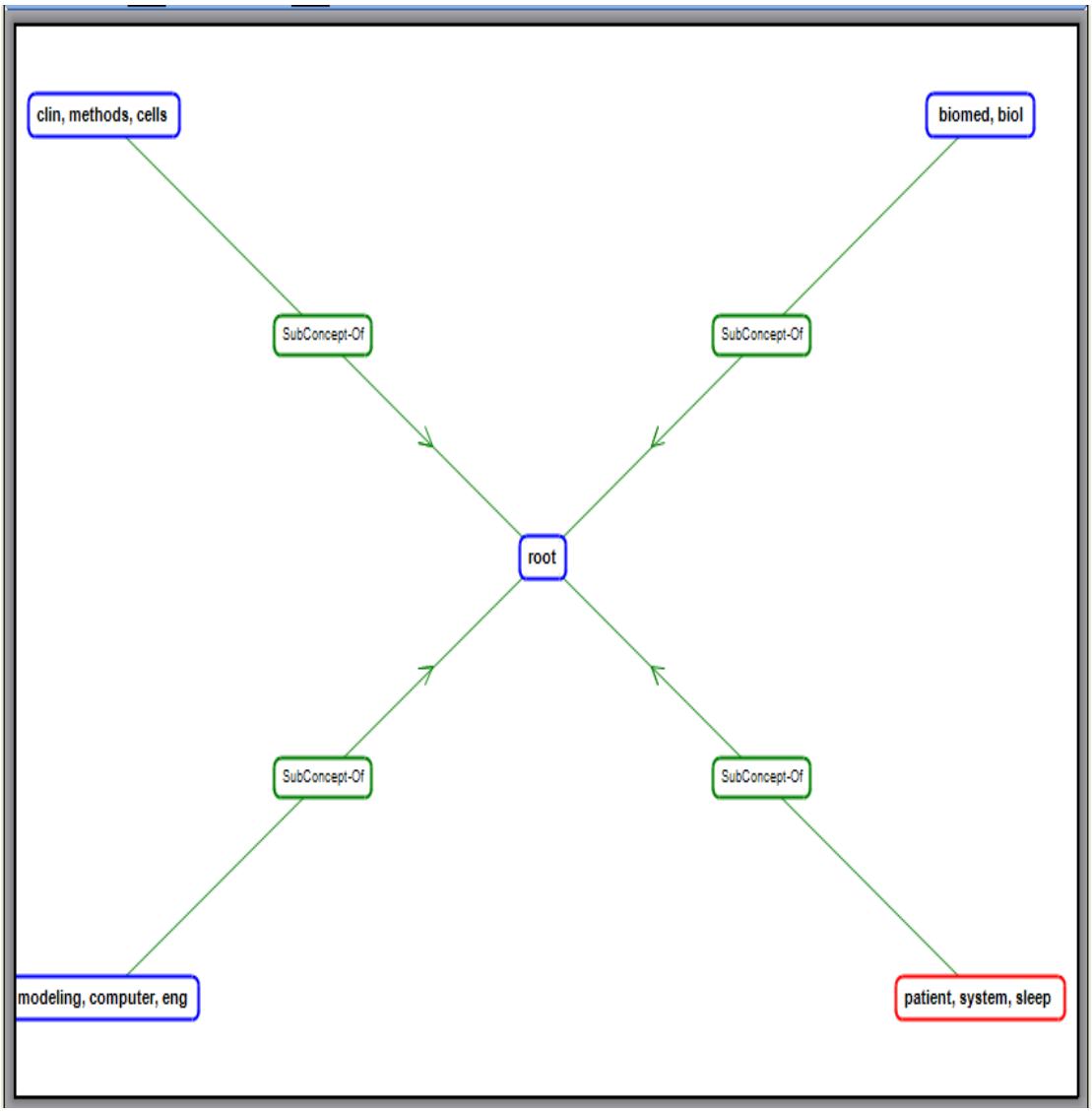


Σχήμα 7.21: Η θεματική οντολογία που δημιουργήσαμε για το πεδίο δεδομένων "Wireless telemedicine emergency, wireless telemedicine ambulance, wireless telemedicine disaster, wireless ambulance, wireless disaster, and wireless emergency".

Εφαρμογή 6η

Σε αυτή την εφαρμογή η οντολογία που δημιουργήσαμε αφορούσε το πεδίο δεδομένων "biosignals and neural networks". Η ανάκτηση των βιβλιογραφικών κειμένων έγινε και πάλι από τη βιβλιοθήκη PUBMED και το μέγεθος της βάσης που δημιουργήσαμε ήταν 205 κείμενα.

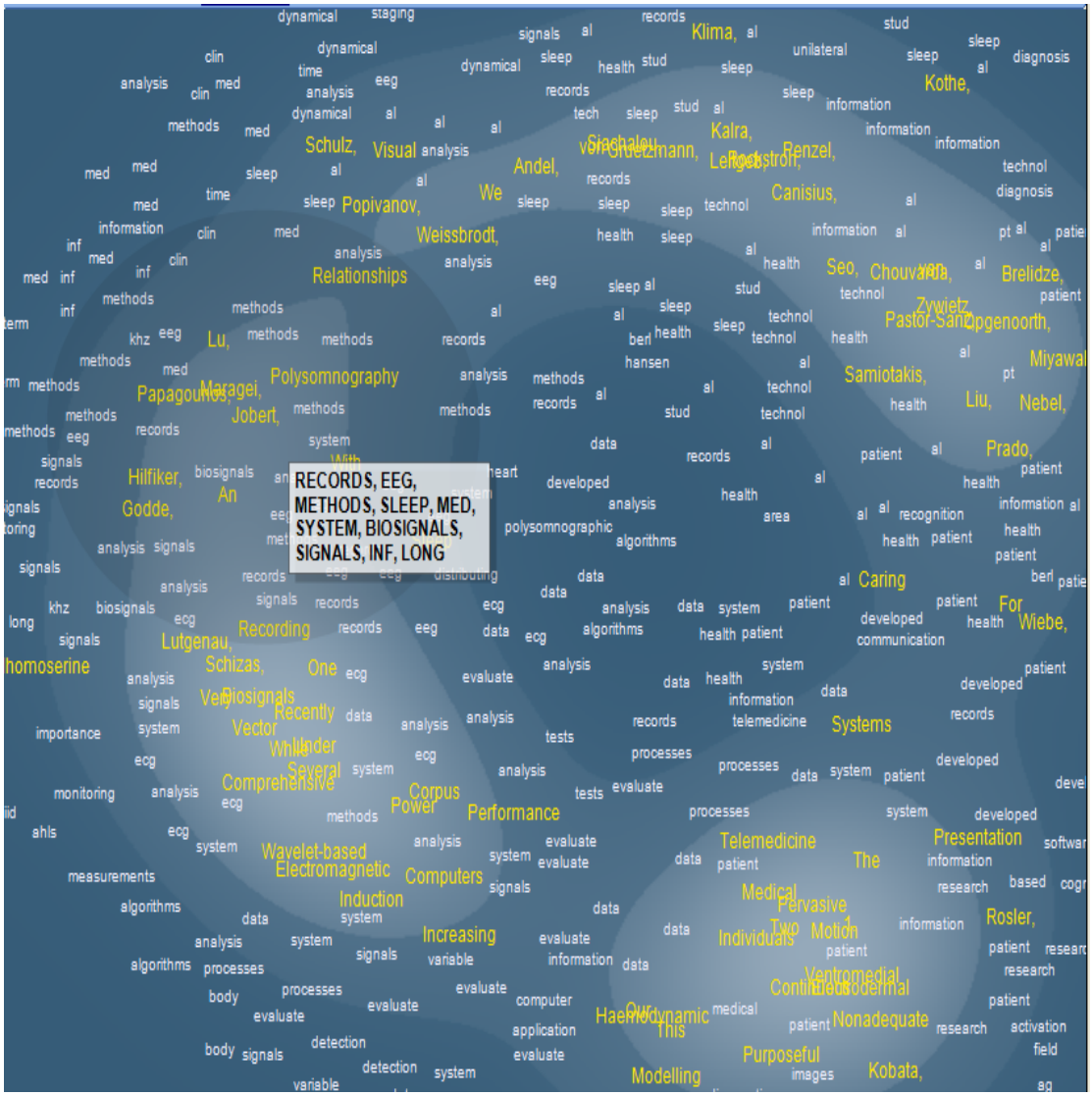
Με τα γνωστά βήματα που αναφέραμε επανειλημμένως, τόσο σε αυτό το κεφάλαιο αλλά και με περισσότερη λεπτομέρεια στο Κεφάλαιο 6 και στο Παράρτημα Α, όπου εξηγούμε βήμα προς βήμα πως δημιουργείται μια οντολογία. Εδώ δείχνω μόνο τις βασικές οθόνες κατά τη διάρκεια της δημιουργίας. Έτσι στο Σχήμα 7.22 είναι η οπτική απεικόνιση της οντολογίας μόλις προσθέσουμε τις πρώτες έννοιες.



Σχήμα 7.22: Η θεματική οντολογία με 4 βασικές έννοιες που μόλις προσθέσαμε.

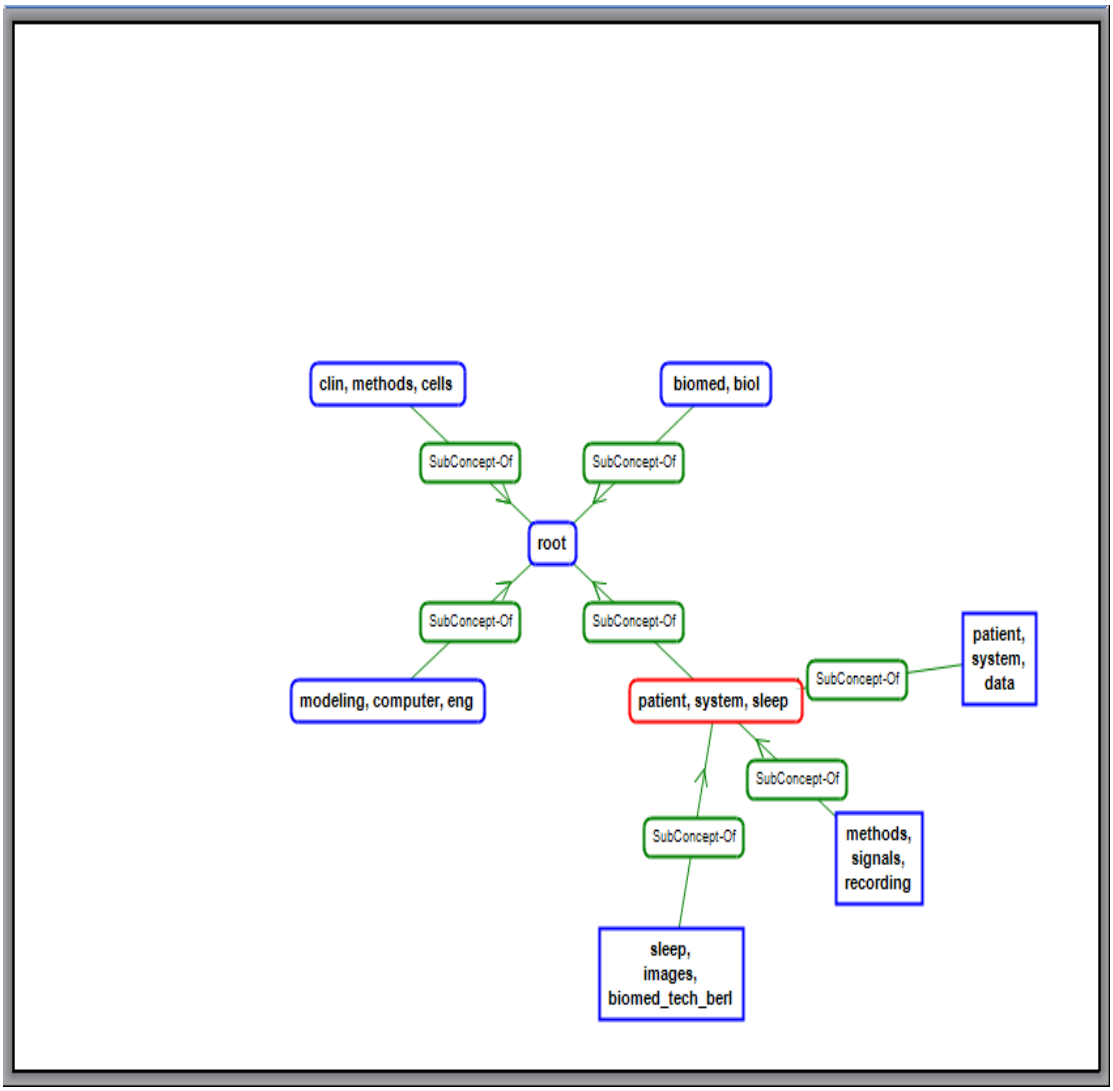
Στη συνέχεια στο Σχήμα 7.23 είναι ο οντολογικός χάρτης για μια έννοια την οποία επιλέγουμε από την οπτική απεικόνιση της έννοιας και η οποία χρωματίζεται με κόκκινο χρώμα όπως φαίνεται και από το Σχήμα 7.77. Ο οντολογικός χάρτης μας

βοηθά να αποφασίσουμε τον αριθμό των εισηγήσεων, όσον αφορά τις παραμέτρους για τους αλγόριθμους ομαδοποίησης. Για παράδειγμα από το Σχήμα 7.23 διακρίνονται τρείς ξεχωριστές ομάδες, γεγονός που μας υποδεικνύει ότι ο αριθμός των εισηγήσεων σε ένα από τους αλγόριθμους ομαδοποίησης που θα επιλέξουμε, θα είναι 3.



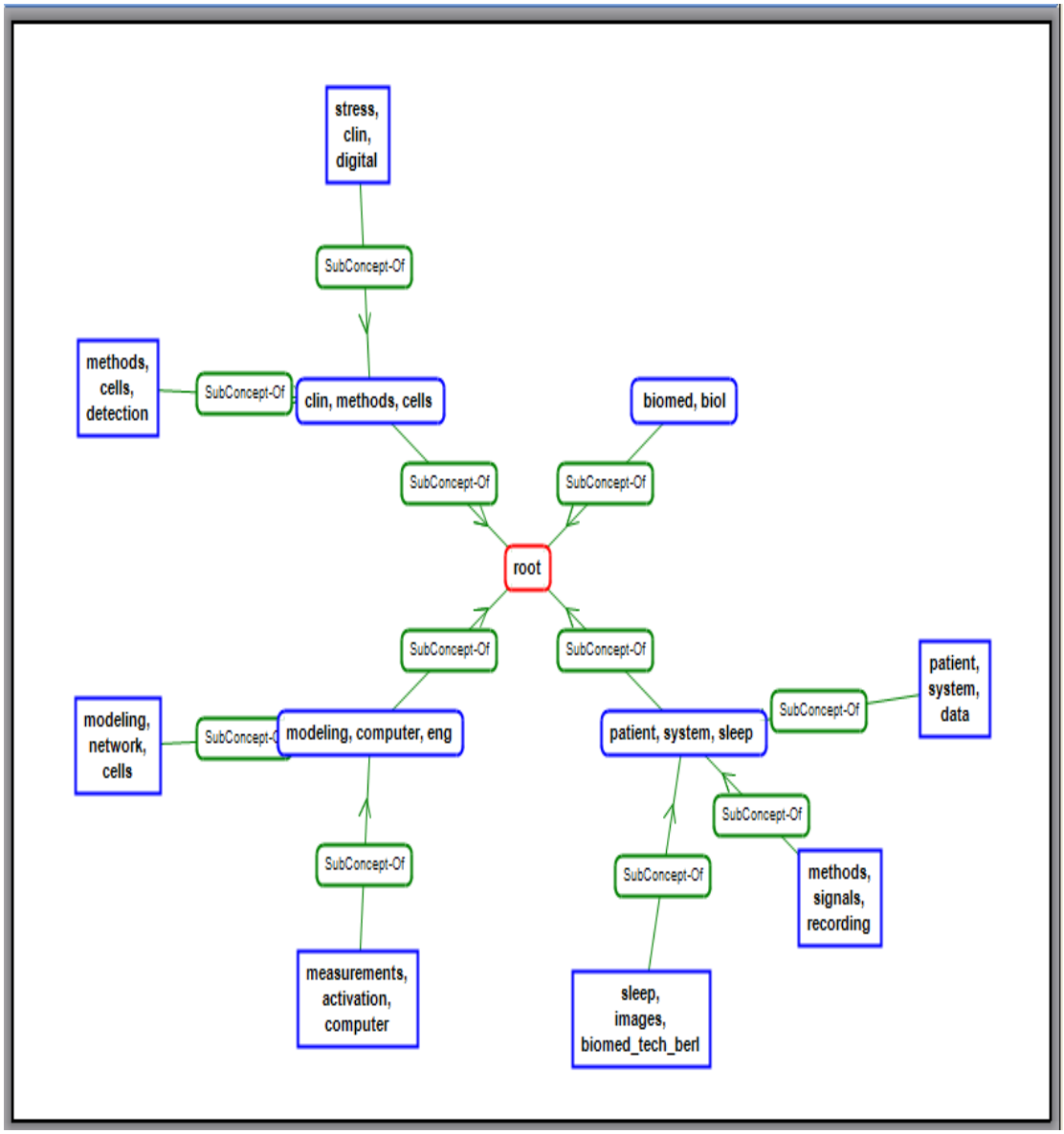
Σχήμα 7.23: Ο οντολογικός χάρτης για την έννοια “patient, system, sleep”.

Αφού λοιπόν εφαρμόσουμε τον αλγόριθμο LSI, με αριθμό εισηγήσεων 3, προκύπτει το Σχήμα 7.24.



Σχήμα 7.24: Οι υπό – έννοιες για την επιλεγμένη έννοια "patient, system, sleep" προστίθενται στην οντολογία.

Στο Σχήμα 7.25 είναι η οντολογία σε ολοκληρωμένη μορφή.



Σχήμα 7.25: Η θεματική οντολογία για το πεδίο δεδομένων “biosignals and neural networks”.

7.4 Συμπεράσματα

Οι πιο πάνω εφαρμογές επιβεβαιώνουν τη δυνατότητα δημιουργίας οντολογιών από το OntoGen από βιοιατρική βιβλιογραφία με τη συστηματική δημιουργία κύριων εννοιών.

Δημιουργώντας οντολογίες από βάσεις δεδομένων με συνδυασμένα έγγραφα, ο στόχος ήταν να ανακαλυφθούν έγγραφα που να έχουν όσο το δυνατό πιο πολλές ομοιότητες. Για αυτό και οι τομές μεταξύ εννοιών σε μια οντολογία υποδεικνύουν κοινά στοιχεία, κάποια σχέση στην οποία πρέπει να κατευθυνθεί η έρευνα μας.

Βασικά μια Οντολογία μπορεί να χρησιμοποιηθεί για την οργάνωση μεγάλων συλλογών από δεδομένα έτσι ώστε οι χρήστες να βρίσκουν πιο εύκολα αυτό που αναζητούν κάθε φορά. Επίσης με αυτές μπορεί να γίνει η περιγραφή ενός πεδίου δεδομένων πιο κατανοητή στους χρήστες. Είναι συνεπώς ένας νέος τύπος δεδομένων.

Οι σημαντικότερες όμως ανακαλύψεις, εξάγονται από τα στατιστικά δεδομένα που παράγει το OntoGen. Τα δεδομένα αυτά όπως ανάφερα μπορούν να οδηγήσουν στην ανακάλυψη χρήσιμων και προηγουμένως άγνωστων πληροφοριών, οι οποίες σχετίζονται με τα φαινόμενα που μελετούνται. Στις εφαρμογές αυτές δεν είδαμε πως γίνεται αυτό, αλλά αναφέρθηκε μόνο και εξηγήθηκε πως γίνεται. Πρόκειται για μια διαδικασία που απαιτεί το ανάλογο γνωστικό υπόβαθρο και ακολουθείται πιστεύω σε έρευνες.

Όλες αυτές οι μεθοδολογίες που είδαμε να ακολουθούνται στις παραπάνω εφαρμογές αναμφισβήτητα αντικαθιστούν όλη τη δουλειά που έκαμνε ένας ερευνητής προηγουμένως χειροκίνητα. Συνεπώς ο χρήστης κερδίζει χρόνο αλλά έχει και καλύτερα αποτελέσματα.

Κεφάλαιο 8

Συμπεράσματα

8.1 Συμπεράσματα

Η εκπόνηση της συγκεκριμένης Ατομικής Διπλωματικής Εργασίας είχε ως πρώτιστο στόχο τη δημιουργία θεματικών οντολογιών σε βιοιατρική βιβλιογραφία και την παραπέρα ανακάλυψη χρήσιμων και προηγουμένως άγνωστων πληροφοριών που σχετίζονται με το θέμα που μελετάται.

Η δημιουργία οντολογιών έγινε με τη βοήθεια του συστήματος OntoGen, το οποίο σύστημα αυτοματοποιεί έως ένα σημείο τη διαδικασία αυτή καθοδηγώντας το χρήστη στις αποφάσεις που παίρνει για τη δημιουργία της οντολογίας. Το σύστημα εισηγείται έννοιες, τις σχέσεις και τα ονόματα τους, αναθέτει αυτόματα στιγμιότυπα στη κάθε έννοια, παρ' όλα αυτά ο χρήστης είναι αυτός που έχει τον όλο χειρισμό του συστήματος με το να μπορεί να ρυθμίζει πλήρως όλες τις ιδιότητες που αφορούν την οντολογία.

Μέσα από τις διάφορες εφαρμογές που έχουμε υλοποιήσει πάνω σε διαφορετικά δεδομένα κάθε φορά, έχει επιτευχθεί η δημιουργία οντολογίας. Η δημιουργία οντολογίας είναι πολύ χρήσιμη για τη συστηματική εξερεύνηση βάσεων δεδομένων. Τα στατιστικά δεδομένα που είδαμε να παράγονται επίσης, μπορούν να οδηγήσουν

στην ανακάλυψη ενδεχομένως σημαντικών πληροφοριών. Σίγουρα όμως η συμμετοχή των εμπειρογνωμόνων είναι κρίσιμη και ουσιαστικότερη για να έχουμε γρηγορότερα και ορθότερα αποτελέσματα.

Συμπερασματικά, το σύστημα OntoGen είναι ένα εργαλείο που θα μπορούσε να χρησιμοποιηθεί από μια μεγάλη μάζα χρηστών, ειδικότερα όμως από ερευνητές. Ερευνητές οι οποίοι καταπιάνονται με οποιοδήποτε θέμα και επιθυμούν μια εις βάθος ανάλυση ή μελετητές ή γιατρούς, οι οποίοι δεν γνωρίζουν πλήρως ένα θέμα. Με το OntoGen, ο ερευνητής εισάγει τη βάση δεδομένων που επιθυμεί να αναλύσει και με τεχνικές εξόρυξης δεδομένων και μηχανική μάθηση, του παρουσιάζονται κάποια πρώτα αποτελέσματα (οντολογικός χάρτης και *.txt.stat file) τα οποία απαιτούν κάποια περαιτέρω ανάλυση για την ανακάλυψη γνώσης.

Γενικά το εργαλείο αξιολογείται ως εύχρηστο και αποδοτικό εργαλείο για την οργάνωση δεδομένων σε μορφή οντολογίας. Σε αξιολόγηση που έγινε για το συγκεκριμένο σύστημα δόθηκε ένας γενικά καλός βαθμός, ο οποίος επιβεβαιώνει τη σωστή διαδρομή της ανάπτυξης που ακολουθήθηκε, αλλά ωστόσο αφήνει περισσότερο από αρκετό χώρο για μελλοντικές βελτιώσεις πάνω στο εργαλείο ώστε να γίνει ακόμα καλύτερο.

8.2 Μελλοντική Εργασία

Στην περαιτέρω ανάπτυξη του συστήματος μπορούν να δημιουργηθούν οι ακόλουθες επεκτάσεις ή βελτιώσεις:

- Η εισαγωγή νέων μεθόδων εξόρυξης κειμένων στο σύστημα. Τέτοια αναβάθμιση όμως του συστήματος μπορεί από τη μια πλευρά να αποφέρει καλύτερα αποτελέσματα, από την άλλη όμως ενέχει ο κίνδυνος το σύστημα να είναι δύσκολο να κατανοηθεί και έτσι να χρησιμοποιείται μόνο από χρήστες που κατέχουν κάποια γνώση στην εξόρυξη δεδομένων.

- Επίσης από τις αξιολογήσεις του συστήματος που έγιναν, στην μελλοντική εργασία των κατασκευαστών του συστήματος είναι να βελτιώσουν τη διεπιφάνεια χρηστών.

- Επέκταση του συστήματος έτσι ώστε να υποστηρίζει συνεργασία χρηστών όταν εκδίδουν μεγαλύτερες οντολογίες και να υποστηρίζει επαναχρησιμοποίηση οντολογιών ή κάποιου μέρους των οντολογιών, οι οποίες παράχθηκαν από άλλους χρήστες.

- Ενσωμάτωση στο σύστημα εργαλείου που να ετοιμάζει τα δεδομένα εισόδου, να φτιάχνει δηλαδή βάσεις δεδομένων και να παρέχονται οι μηχανισμοί για οποιαδήποτε προ-επεξεργασία στα δεδομένα αν απαιτείται.

- Ακόμα μια πιθανή κατεύθυνση μελλοντικής εργασίας θα ήταν η όλη διαδικασία να γινόταν πιο αυτοματοποιημένη και να μειωθεί η ανάγκη για αλληλεπίδραση με το χρήστη.

Βιβλιογραφία

[1] Ramez Elmasri, Shamkant B. Navathe, "Fundamentals of Database System", Third Edition, ADDISON-WESLEY, 1999.

[2] B. Fortuna, M. Grobelnik, D. Mladenic, "OntoGen: Semi-automatic Ontology Editor", HCI 2007:309-318.

[3] B. Fortuna, M. Grobelnik, D. Mladenic, "Semi-automatic Data-driven Ontology Construction System" Proceedings of the 9th International multi-conference Information Society IS-2006, Ljubljana, Slovenia.

[4] B. Fortuna, M. Grobelnik, D. Mladenic, "System for Semi-automatic Ontology Construction". Demo at ESWC 2006, June 11-14, 2006, Budva, Montenegro.

[5] B. Fortuna, M. Grobelnik, D. Mladenic, "Background knowledge for ontology Construction", WWW 2006:949-950.

[6] B. Fortuna, D. Mladenic, M. Grobelnik, "Visualization of Text Document Corpus", Informatica (Slovenia) (INFORMATICASI) 29(4):497-504 (2005).

[7] B. Fortuna, M. Grobelnik, D. Mladenic, "Semi-automatic Construction of Topic Ontologies", EWMF/KDO 2005:121-131.

[8] J. Y. Bansard, D. Rebholz-Schuhmann, G. Cameron, D. Cameron, D. Clark, E. van Mulligen, F. Beltrame, E. Del Hoyo Barbolla, F. Martin-Sanchez, LMilanesi, I.Tollis, J. van der Lei, and J. L Coatrieux, "Medical Informatics and Bioinformatics: A Bibliometric Study", IEEE Transactions on Information Technology in Biomedicine (TITB) 11(3):237-243 (2007).

[9] Sam Scott, "Some New Evidence for Concept Stability", Department of Cognitive Science, Carletor University, Ottawa.

[10] Asunción Gómez-Pérez,V. Richard Benjamins, “Applications of Ontologies and Problem-Solving Methods”, AI Magazine 20(1): 119-122 (1999).

[11] B. Mathiak, S. Eckstein, “Five Steps to Text Mining in Biomedical Literature”, Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics.

[12] B. Fortuna, N. Lavrac, P. Velardi, “Advancing Topic Ontology Learning Through Term Extraction”.

[13] http://ontogen.ijs.si/

[14] http://kt.ijs.si/Dunja/textgarden/

Παράρτημα Α

Σύστημα Δημιουργίας Οντολογιών OntoGen
Εγχειρίδιο Χρήσης

Πίνακας Περιεχομένων

Α.1 Εγκατάσταση

Α.2 Χρησιμοποιώντας το Σύστημα

A.1

Εγκατάσταση

Εισαγωγή

Το εγχειρίδιο αυτό προορίζεται για τους χρήστες τους συστήματος OntoGen το οποίο διατίθεται δωρεάν στο διαδίκτυο και σχεδιάστηκε και κατασκευάστηκε από τους Blaz Fortuna, Marko Grobelnik και Dunja Mladenic. Το σύστημα είναι ένας ημι – αυτόματος και καθοδηγούμενος από δεδομένα εκδότης, που επικεντρώνεται στη δημιουργία θεματικών οντολογιών. Επίσης συνδυάζει τεχνικές εξόρυξης κειμένου με μια αποδοτική διεπιφάνεια, έτσι ώστε να μειώσει το χρόνο που ξοδεύεται αλλά και τη πολυπλοκότητα για το χρήστη.

Στόχος του κεφαλαίου αυτού είναι να οδηγήσει βήμα προς βήμα στην εγκατάσταση του συστήματος.

1.1 Πριν Αρχίσετε

Πριν αρχίσετε να εκτελείτε τα βήματα εγκατάστασης του συστήματος θα πρέπει να:

1. Βεβαιωθείτε ότι έχετε μαζί σας το εγχειρίδιο χρήσης του λειτουργικού συστήματος που χρησιμοποιείται.
2. Πηγαίνετε στην ιστοσελίδα http://ontogen.ijs.si/?page_id=10 , όπου από εκεί θα γίνει η εγκατάσταση του εργαλείου OntoGen.

1.2 Απαιτήσεις Συστήματος

Για τη χρήση του συστήματος υπάρχουν οι ακόλουθες απαιτήσεις σε λογισμικό

- .NET framework 2.0

1.3 Εγκατάσταση και Προετοιμασία

Ακολουθήστε τα πιο κάτω βήματα για την εγκατάσταση και διαμόρφωση του συστήματος:

1. Δημιουργία της Βάσης Δεδομένων χρησιμοποιώντας το εγχειρίδιο χρήσης του λογισμικού EndNote X.
2. Φόρτωμα της Βάσης Δεδομένων στο σύστημα OntoGen.

A.2

Χρησιμοποιώντας το Σύστημα

Εισαγωγή

2.1 Δημιουργία και Αποθήκευση Οντολογιών

2.2 Βασική Διαχείριση Εννοιών

2.3 Εισήγηση και Πρόσθεση Εννοιών

2.4 Οπτική Απεικόνιση Οντολογίας

2.5 Διαχείριση εγγράφων των εννοιών

2.6 Οπτική Απεικόνιση Εννοιών

2.7 Διαχείριση Σχέσεων

2.8 Ενημέρωση Υπάρχουσας Οντολογίας

Εισαγωγή

Σκοπός του κεφαλαίου αυτού είναι η παρουσίαση και η ανάλυση τρόπου χρήσης των διαφόρων λειτουργιών του διαχειριστικού μέρους του συστήματος.
Οι λειτουργίες αυτές αφορούν κυρίως:

1. Εισήγηση Εννοιών
2. Πρόσθεση εννοιών με διάφορους τρόπους
3. Απεικόνιση οντολογίας
4. Απεικόνιση εννοιών
5. Διαχείριση εγγράφων

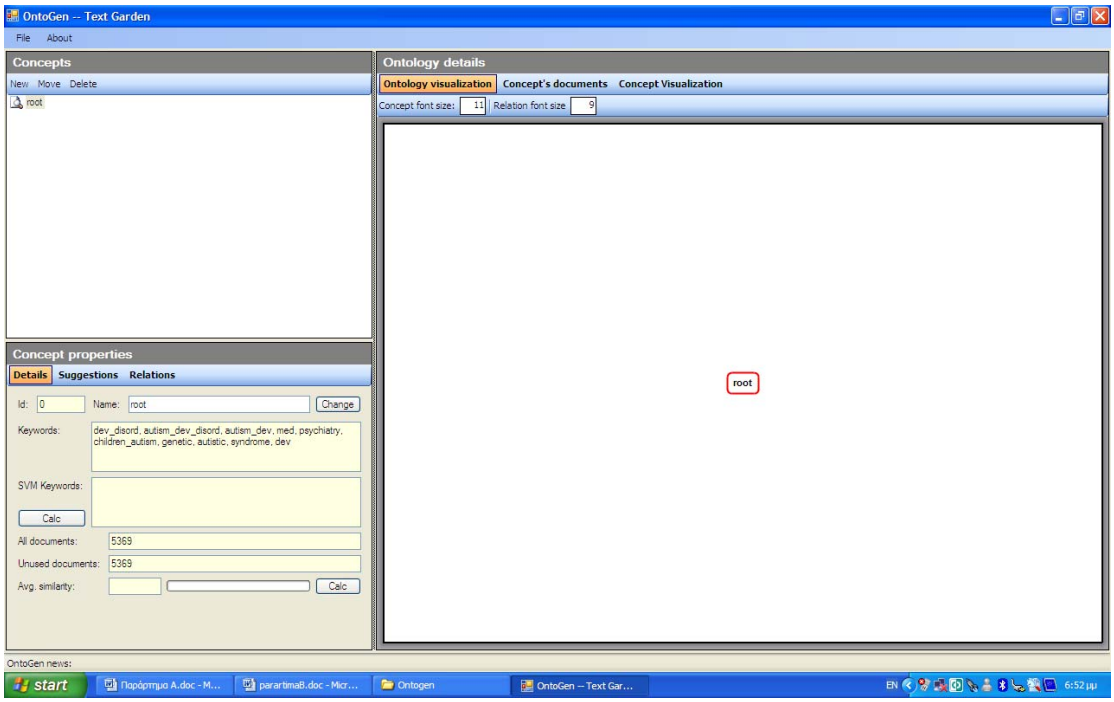
2.1 Δημιουργία και Αποθήκευση Οντολογιών

Αφού λοιπόν έχει γίνει το φόρτωμα της βάσης δεδομένων στο σύστημα, η οποία γίνεται μέσω της επιλογής :

File> Folder or Named Line-Documents *or* ***Bag-of-Words*** *or* ***Bag-of-Words with Weights***

Η επιλογή γίνεται αναλόγως σε ποια μορφή έχουμε τα κείμενα μας φυλαγμένα στη βάση μας. Η επιλογή Folder είναι για HTML ή για Text files, τα οποία είναι φυλαγμένα σε κάποιο folder. Η επιλογή Named Line-Documents φορτώνει έγγραφα από ένα αρχείο όπου κάθε γραμμή μεταφράζεται σαν ένα έγγραφο με την πρώτη λέξη στη γραμμή να είναι ο τίτλος του εγγράφου.

Αφού λοιπόν φορτωθεί η βάση, προτού κάμει οτιδήποτε ο χρήστης εμφανίζεται αυτόματα η πρώτη έννοια στον οντολογικό χάρτη (root concept) και δίπλα στο παράθυρο της διαχείρισης των εννοιών μπορούμε να δούμε τα κύρια κλειδιά (keywords) που περιγράφουν αυτή την έννοια. Επίσης στο πάνω αριστερό παράθυρο υπάρχει μόνο η έννοια root. Αυτό που περιγράφεται φαίνεται πιο κάτω στην Εικόνα1. Η έννοια root περιέχει ακόμα όλα τα έγγραφα που υπάρχουν στη βάση, αφού ακόμα δεν έχει γίνει καμία επεξεργασία, δεν έχει εφαρμοστεί κανένας αλγόριθμος κατηγοριοποίησης.



Εικόνα 1: Φόρτωμα βάσης δεδομένων και δημιουργία βασικής έννοιας

Ο χρήστης μπορεί να αποθηκεύσει τις οντολογίες που θα δημιουργηθούν με το εργαλείο OntoGen σαν Proton Topic Ontology, RDF schema ή OWL ontology.

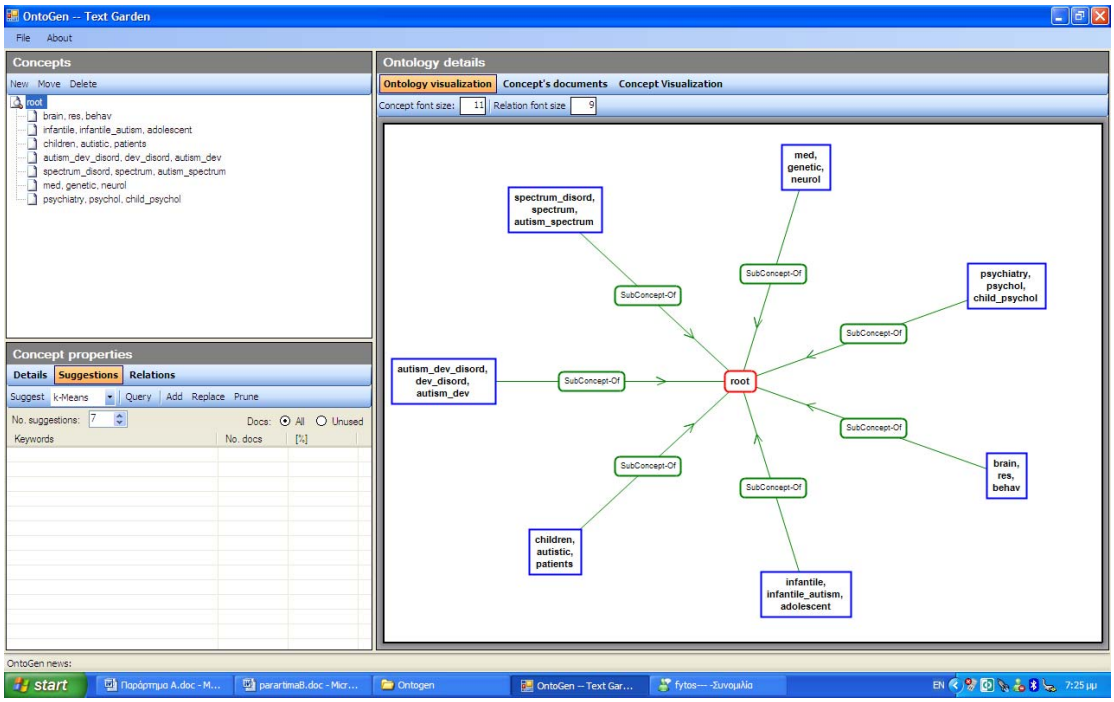
Για την περαιτέρω δημιουργία οντολογίας θα δούμε στη συνέχεια στις διάφορες υπό – ενότητες του εγχειριδίου.

2.2 Βασική Διαχείριση Εννοιών

Ο χρήστης έχει την ευχέρεια μέσα από το αριστερό πάνω παράθυρο να έχει μια γρήγορη άποψη όλων των εννοιών και της θέσης τους μέσα στην εννοιολογική ιεραρχία αλλά επίσης μπορεί να ασκεί έλεγχο πάνω στις έννοιες όσον αφορά και τη θέση τους.

Εκτός από τη δημιουργία καινούργιας έννοιας, μπορεί ακόμα να επανατοποθετεί μια υπάρχουσα έννοια σε μια διαφορετική θέση στην ιεραρχία οντοτήτων ή να την αφαιρέσει εντελώς από την οντολογία. Αυτό γίνεται με το να επιλέξει πρώτα ο χρήστης την έννοια από το παράθυρο “Concepts” στην εννοιολογική ιεραρχία (πάνω αριστερά) και στη συνέχεια να επιλέξει το κουμπί με την ανάλογη πράξη που

επιθυμεί να εκτελεστεί: “New”, “Move” , ή “Delete”. Αν ο χρήστης μετακινεί κάποια έννοια τότε εμφανίζεται ένα παράθυρο που το ρωτάει να επιλέξει την έννοια προορισμού.

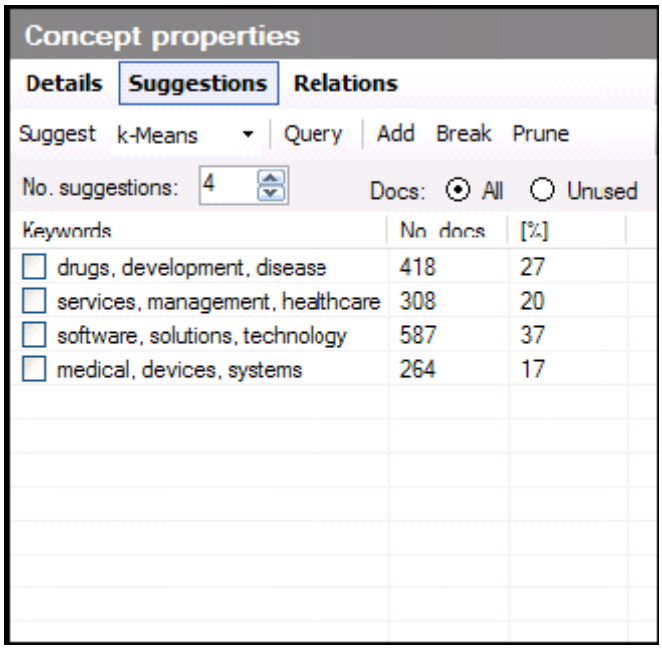


Εικόνα 2: Στο παράθυρο πάνω αριστερά βλέπουμε την εννοιολογική ιεραρχία και τις τρείς διαθέσιμες επιλογές “New”, “Move”, “Delete”

2.3 Εισήγηση και Πρόσθεση Εννοιών

Ένα από τα κυριότερα μέρη του συστήματος είναι η μάθηση εννοιών, όπου στο σύστημα OntoGen αυτή η λειτουργικότητα είναι προσβάσιμη μέσω του tab “Concept – suggestion”.

Ο χρήστης μέσω του παράθυρου “Concept properties” κάτω από το tab “Suggestions” μπορεί να εφαρμόσει τους δύο αλγόριθμους που του διατίθενται από το σύστημα για εισήγηση εννοιών. Ο χρήστης έχει στη διάθεση του λοιπόν τους αλγόριθμους “LSI” και “k-means” (βλέπε Εικόνα 3).

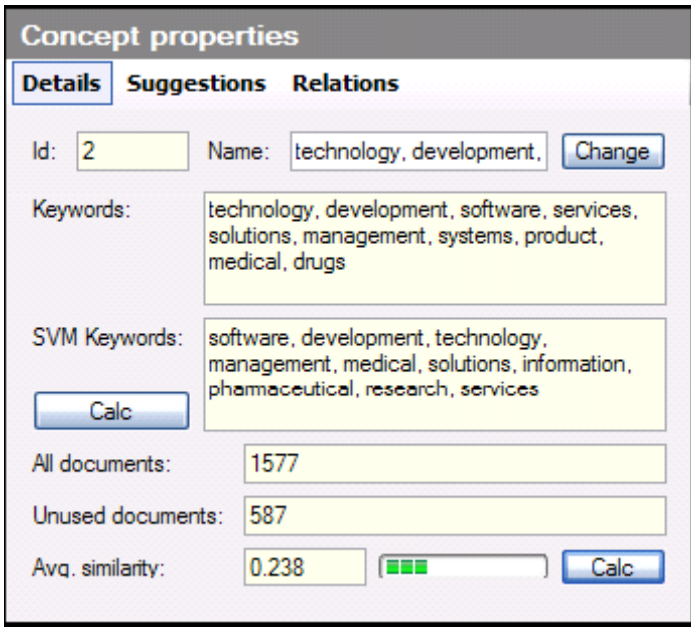


Εικόνα 3: Λίστα εισηγημένων εννοιών

Από τη λίστα των εισηγημένων εννοιών ο χρήστης μπορεί να επιλέξει ποιες από αυτές επιθυμεί να προσθέσει στην οντολογία που κατασκευάζει, στον οντολογικό χάρτη, απλώς επιλέγοντας τις και στη συνέχεια πατώντας "Add".

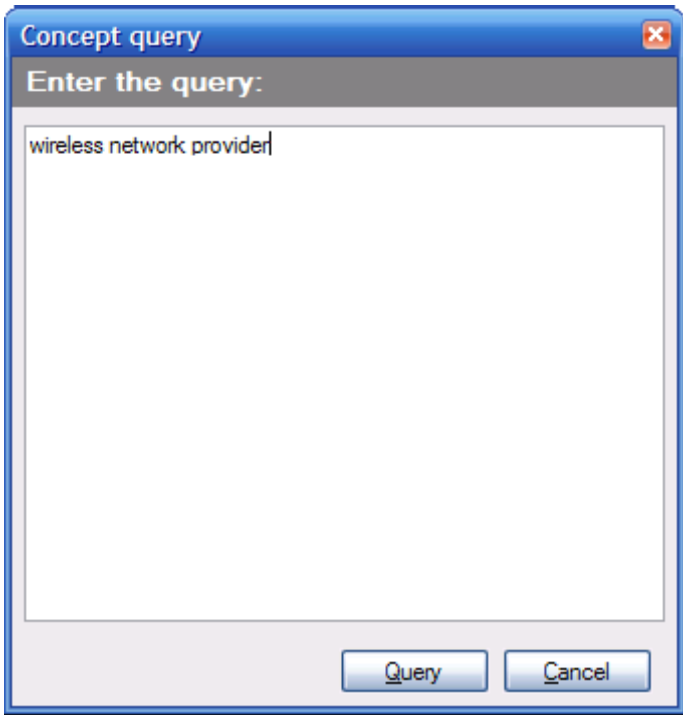
Σε κάθε έννοια ο χρήστης μπορεί να προσθέσει υπό – έννοια/ες με τον ίδιο τρόπο που αναφέραμε πιο πάνω, απλώς επιλέγοντας την από τον χάρτη. Τότε η επιλεγμένη έννοια χρωματίζεται με κόκκινο χρώμα.

Όταν μια έννοια ή υπό – έννοια επιλεχθεί τότε ταυτόχρονα μπορούμε να δούμε ότι την αφορά, ότι σχετίζεται με αυτή. Συγκεκριμένα, στο παράθυρο "Concept properties" κάτω από το tab "Details", έχουμε την ευχέρεια να δούμε το Id number της, το όνομα της και ταυτόχρονα να τη μετονομάσουμε αν επιθυμούμε, τις λέξεις κλειδιά που την περιγράφουν τα οποία υπολογίζονται και εμφανίζονται με δύο τρόπους στο παράθυρο κάτω από τις λέξεις keywords και SVM keywords. Επίσης αναφέρει πόσα στιγμιότυπα, δηλαδή πόσα έγγραφα περιλαμβάνει η έννοια αυτή (All documents), τον αριθμό των στιγμιότυπων που δεν βρίσκονται σε κάποιες από τις υπό – έννοιες (Unused documents) και το μέσο όρο ομοιότητας μέσα στην έννοια (Avg. Similarity). Βλέπε Εικόνα 4.



Εικόνα 4: Λεπτομέρειες που αφορούν την επιλεγμένη έννοια

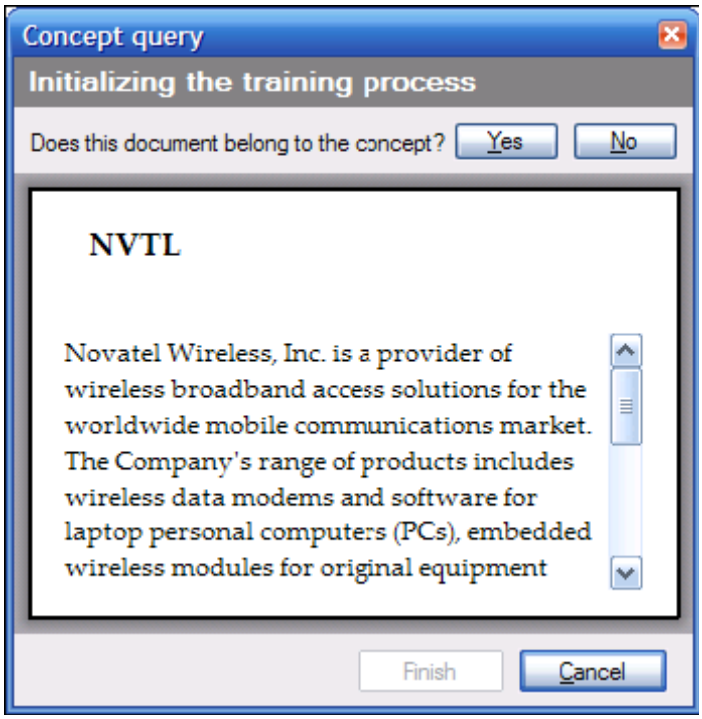
Μια άλλη μέθοδος για πρόσθεση εννοιών, εκτός από τις δύο που είδαμε πιο πριν που γίνονται με τη χρήση των δύο αλγόριθμων, το σύστημα διαθέτει ακόμα ένα τρόπο. Αυτός ο τρόπος είναι μέσω "Query" όπως θα το δείτε στο παράθυρο "Concept properties" κάτω από το tab "Suggestions". Μέσω λοιπόν αυτής της παροχής με το που πατά ο χρήστης το κουμπί "Query" ανοίγεται ένα παράθυρο το οποίο ζητά από το χρήστη να εισάγει ένα query, μια πρόταση δηλαδή, ή έστω κάποιες λέξεις κλειδιά που πιστεύουμε ότι θα αποτελέσουν την επόμενη υπό – έννοια ή έννοια που πρέπει να προστεθεί στο χάρτη της οντολογίας μας. Άρα αυτό που γράφουμε πρέπει να έχει σχέση με το αντικείμενο που μελετάμε, με τη βάση δεδομένων που επεξεργαζόμαστε. Αφού εισάγει το "query" ο χρήστης στο χώρο που του υποδεικνύεται έπειτα πατάει το κουμπί "Query" για να ξεκινήσει η διαδικασία (βλέπε Εικόνα 5).



Εικόνα 5: Εισαγωγή query

Με το που πατάει λοιπόν ο χρήστης το κουμπί "Query", ξεκινά μια διαδικασία ερωτήσεων Yes ή No εάν το συγκεκριμένο έγγραφο (στιγμιότυπο) που εμφανίζεται μπροστά στο χρήστη κάθε φορά, ανήκει στην συγκεκριμένη έννοια. Ο χρήστης αρχικά βλέποντας με μια πρώτη ματιά τα έγγραφα που του παρουσιάζονται μπορεί να απαντήσει με Yes ή No (βλέπε Εικόνα 6).

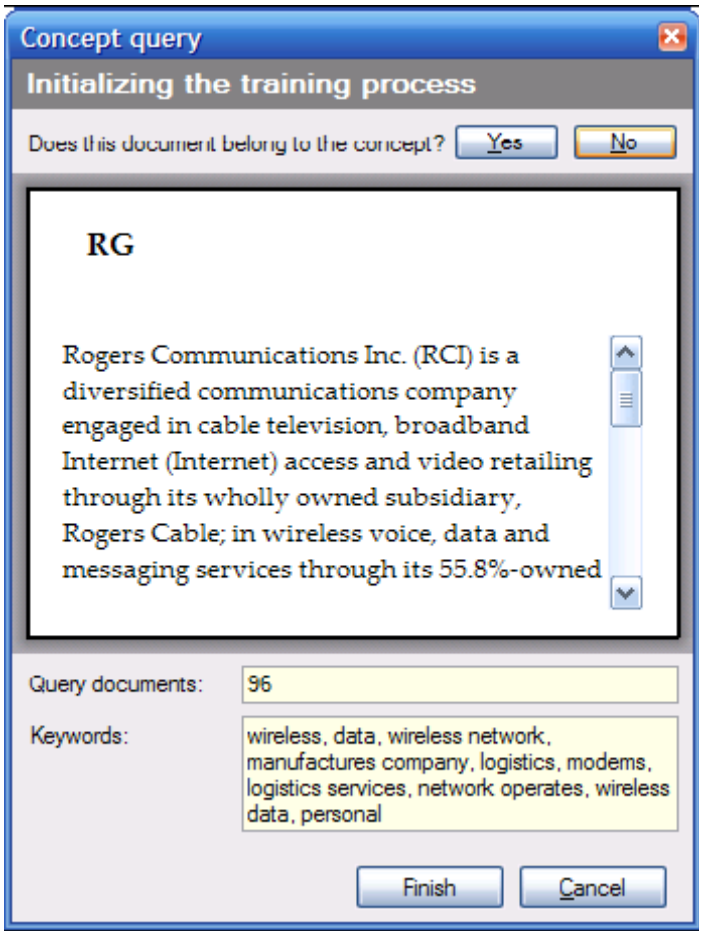
Μετά όμως, όταν θα παρθούν κάποια αρχικά δείγματα από το χρήστη και το σύστημα έχει προσδιορίσει στο περίπου ποια έννοια επιθυμεί ο χρήστης να προσθέσει, το σύστημα τότε εμφανίζει μερικές επιπρόσθετες πληροφορίες σχετικές με την έννοια. Όπως, το μέγεθος μέχρις στιγμής (δηλ. τον αριθμό των εγγράφων που κατηγοριοποιήθηκαν στην έννοια) και τις πιο σημαντικές λέξεις κλειδιά για την έννοια (βλ. Εικόνα 7).



Εικόνα 6: Το σύστημα ρωτά το χρήστη αν το συγκεκριμένο στιγμιότυπο ανήκει ή όχι στην έννοια.

Ο χρήστης μπορεί να συνεχίσει να απαντά στις ερωτήσεις Yes ή No, ή να τερματίσει αυτή τη διαδικασία πατώντας "Finish". Όσες πιο πολλές ερωτήσεις απαντήσει ο χρήστης, τόσο πιο σωστή θα είναι η ανάθεση των εγγράφων (στιγμιότυπων) στην τελική έννοια που θα σχηματιστεί.

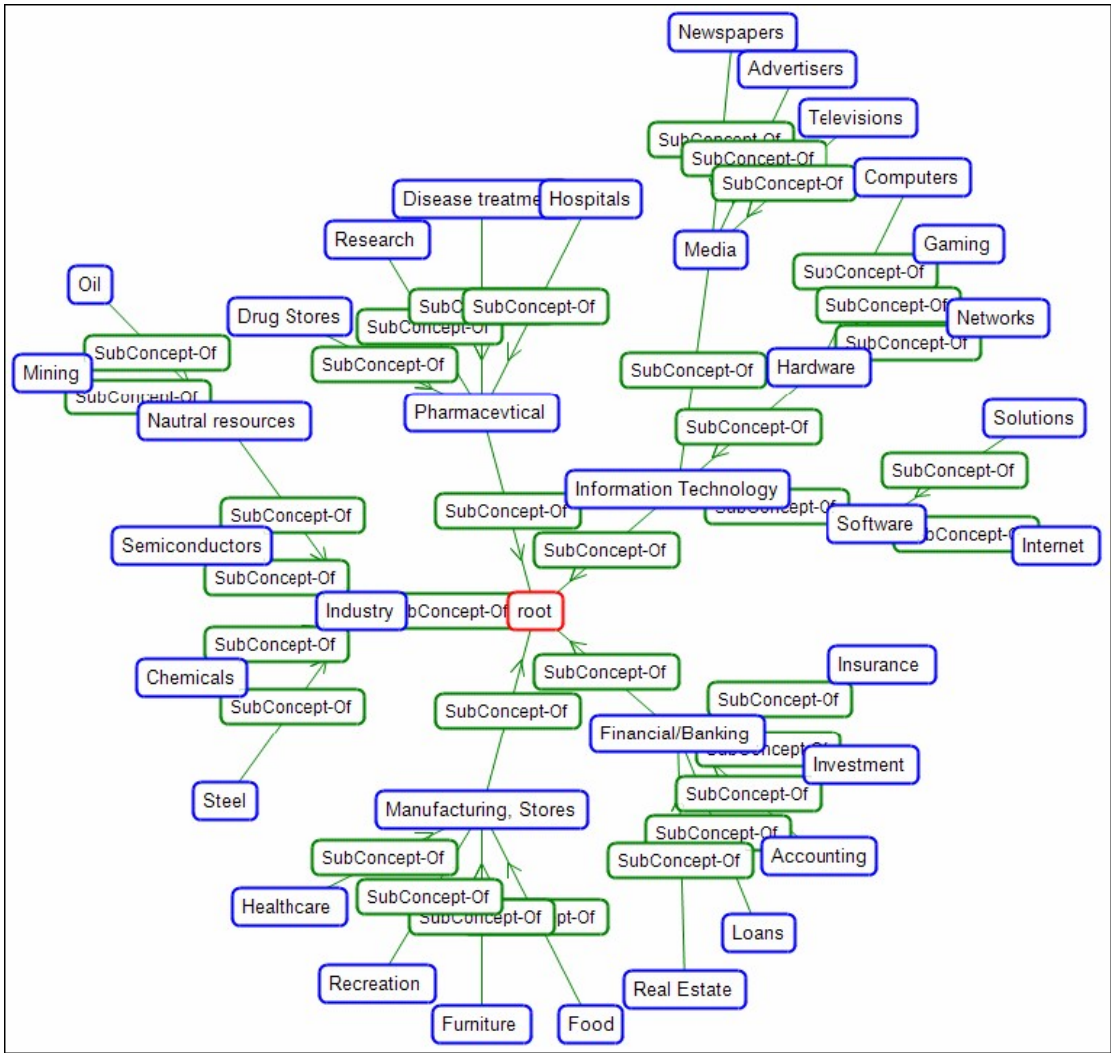
Πατώντας "Finish" η καινούργια έννοια ή υπό – έννοια προστίθεται στην οντολογία. Αν είχαμε επιλέξει προηγουμένως κάποια έννοια στον οντολογικό χάρτη, τότε προστίθεται ως υπό – έννοια της επιλεγμένης έννοιας.



Εικόνα 7: Όταν κάποια αρχικά δείγματα έχουν μαζευτεί από το χρήστη, το σύστημα εμφανίζει κάποιες επιπρόσθετες πληροφορίες για την έννοια που επιθυμεί να προσθέσει ο χρήστης.

2.4 Οπτική Απεικόνιση Οντολογίας

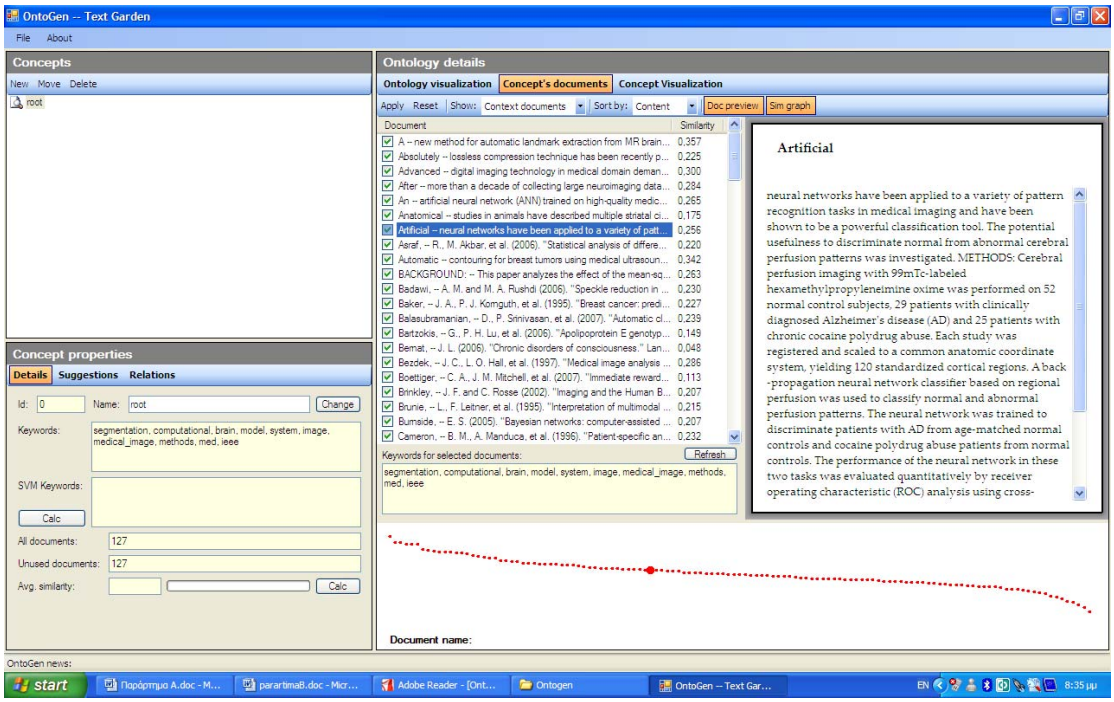
Ο χρήστης έχει και την ευχέρεια να βλέπει τι κατασκευάζει καθ' όλη τη διάρκεια της δημιουργίας της οντολογίας στη δεξιά μεριά του κυρίως παράθυρου (Ontology visualization). Έχουμε τη root concept να εμφανίζεται πάντα στο κέντρο και στη συνέχεια ότι άλλες έννοιες, υπό – έννοιες προσθέσουμε να τοποθετούνται σε κύκλο γύρω από αυτήν. Ο χρήστης μπορεί να επιλέξει μια έννοια από την οντολογία, απλώς επιλέγοντας την. Η επιλεγμένη έννοια χρωματίζεται κόκκινη, ενώ οι υπόλοιπες είναι μπλε και οι σχέσεις είναι με πράσινο χρώμα (Βλέπετε Εικόνα 8).



Εικόνα 8: Οπτική απεικόνιση οντολογίας

2.5 Διαχείριση εγγράφων των εννοιών

Κάτω από το tab "Concept Documents", ο χρήστης έχει την ευχέρεια να δει όλα τα έγραφα που περιέχονται στη βάση μας ή που αναφέρονται σε μια έννοια που επιλέγουμε με τη μορφή λίστας. Δίπλα από τη λίστα των εγγράφων αυτών (στιγμιότυπων), μπορούμε να δούμε το περιεχόμενο ενός εγγράφου αν το επιλέξουμε από τη λίστα. Στο κάτω μέρος του ίδιου παραθύρου υπάρχει η γραφική ομοιότητας. Εικόνα 9.

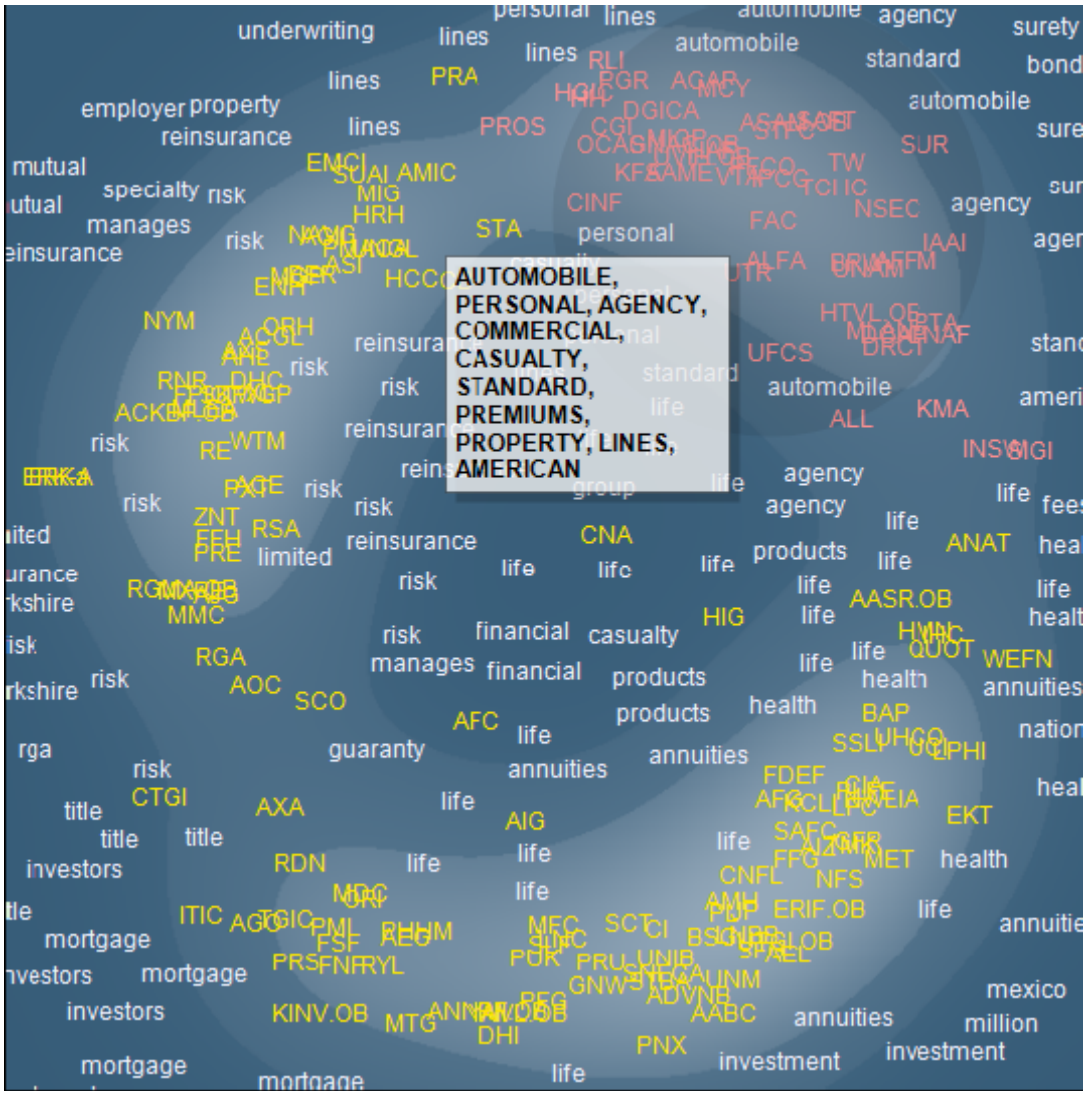


Εικόνα 9: Στην εικόνα πιο πάνω φαίνεται το επιλεγμένο έγγραφο, δίπλα το περιεχόμενο του και κάτω η γραφική παράσταση ομοιότητας.

2.6 Οπτική Απεικόνιση Εννοιών

Κάτω από το tab "Concept Visualization", γίνεται μια οπτική απεικόνιση των εννοιών, όπως φαίνεται στην Εικόνα 10.

Στην απεικόνιση αυτή έχουμε τις έννοιες να εμφανίζονται κατά ομάδες. Μετακινώντας ο χρήστης το ποντίκι του στην οθόνη μπορεί να δει τις πιο κοινές λέξεις κλειδιά για κάθε περιοχή του χάρτη.

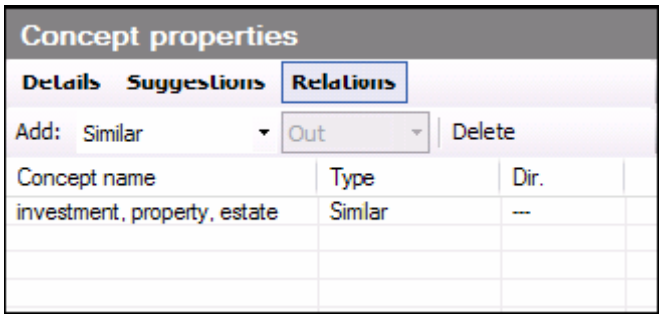


Εικόνα 10: Στο χάρτη διακρίνονται τρεις ομάδες και τα στιγμιότυπα που είναι με ροζ χρώμα, είναι έτσι διότι τα επέλεξε ο χρήστης ούτως ώστε να τα προσθέσει ως υπό – έννοια στην οντολογία.

2.7 Διαχείριση Σχέσεων

Ο χρήστης μπορεί να ορίσει δικούς του τύπους σχέσεων εκτός της σχέσης "sub-concept" που ήδη υπάρχει.

Αυτή η λειτουργικότητα παρέχεται μέσω του παράθυρου κάτω αριστερά στην οθόνη του συστήματος, κάτω από το tab "Relation" όπως φαίνεται και στην Εικόνα 11.

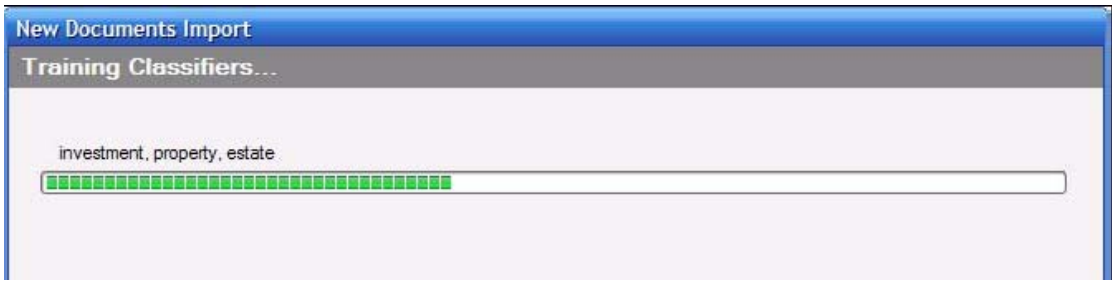


Εικόνα 11: Για ορισμό σχέσεων στην οντολογία

2.8 Ενημέρωση Υπάρχουσας Οντολογίας

Αν ο χρήστης έχει ήδη δημιουργήσει την οντολογία του όμως μετά από κάποιο χρονικό διάστημα επιθυμεί να την ενημερώσει με νέα έγγραφα (στιγμιότυπα), τότε έχει αυτή τη δυνατότητα χωρίς να χρειάζεται να ξαναρχίσει να κτίζει την οντολογία από την αρχή.

Μέσω του μενού, επιλέγει: ***File>Include*** και το σύστημα μπορεί να φορτώσει τα νέα στιγμιότυπα από το ***Folder*** ή ***Named-Lined Documents***. Το σύστημα στη συνέχεια κατηγοριοποιεί τα νέα δεδομένα στην οντολογία (Εικόνα 12).



Εικόνα 12: Το σύστημα εκπαιδεύει τους SVM ταξινομητές στα νέα δεδομένα που μόλις φορτώθηκαν και τους χρησιμοποιεί για την ταξινόμηση νέων στιγμιότυπων.

Παράρτημα Β

Σύστημα Δημιουργίας Βάσης Δεδομένων EndNote X
Εγχειρίδιο Χρήσης

Πίνακας Περιεχομένων

Β.1 Εγκατάσταση

Β.2 Χρησιμοποιώντας το Σύστημα

B.1

Εγκατάσταση

Εισαγωγή

Το εγχειρίδιο αυτό παραθέτεται με σκοπό να μας βοηθήσει στο να φτιάξουμε εύκολα και γρήγορα βάσεις δεδομένων οι οποίες θα αποτελέσουν την είσοδο δεδομένων για το σύστημα που μελετάμε, το σύστημα OntoGen.

Το EndNote X είναι ένα online εργαλείο αναζήτησης το οποίο παρέχει ένα απλό τρόπο να ψάχνουμε online και να γίνεται η εξόρυξη των αναφορών απευθείας στο EndNote. (Το EndNote μπορεί να ενημερώσει επίσης αρχεία δεδομένων, CD – ROMs και βιβλιοθήκες βάσεων δεδομένων).

Στόχος λοιπόν αυτού του κεφαλαίου είναι να οδηγήσει βήμα προς βήμα στην εγκατάσταση του λογισμικού αυτού εργαλείου και να απαριθμήσει τις απαιτήσεις στου συστήματος.

1.1 Πριν Αρχίσετε

Πριν αρχίσετε να εκτελείται τα βήματα εγκατάστασης του συστήματος θα πρέπει να:

1. Βεβαιωθείτε ότι έχετε μαζί σας το CD-ROM εγκατάστασης του παρόντος λογισμικού.

2. Βεβαιωθείτε ότι έχετε μαζί σας το serial number για την εγκατάσταση του προγράμματος.

1.2 Απαιτήσεις Συστήματος

Για τη χρήση του συστήματος υπάρχουν οι ακόλουθες απαιτήσεις σε υλικό:

- Ηλεκτρονικός Υπολογιστής Pentium 450 MHz ή με γρηγορότερο επεξεργαστή
- Τουλάχιστον 256 MB RAM
- Τουλάχιστον 180 MB ελεύθερος δίσκος
- Με σκοπό να χρησιμοποιηθεί η εντολή του EndNote "Connect", έτσι ώστε να γίνει αναζήτηση σε κάποια απομακρυσμένη βάση δεδομένων, απαιτείται σύνδεση στο διαδίκτυο.

Επίσης απαιτείται το πιο κάτω λογισμικό:

- Windows 2000, με Service Pack 3 ή πιο νέο
- Windows XP

1.3 Εγκατάσταση και Προετοιμασία

Για την εγκατάσταση και διαμόρφωση του συστήματος ακολουθήστε τα πιο κάτω βήματα:

1. Ξεκινήστε το πρόγραμμα εγκατάστασης EndNote, αφού έχετε σιγουρευτεί ότι καμιά άλλη εφαρμογή δεν τρέχει.

2. Εάν έχετε κατεβάσει τον EndNote installer, τότε πατήστε δύο φορές στο installer file για να ξεκινήσει το πρόγραμμα εκκίνησης του EndNote.
   Εάν όμως έχετε το CD, τότε εισάγετε το CD στο CD-ROM drive και το πρόγραμμα εκκίνησης θα ξεκινήσει.

3. Ακολουθήστε τις οδηγίες που θα εμφανιστούν στην οθόνη για να ολοκληρωθεί η εγκατάσταση. Χρησιμοποιήστε το κουμπί "Next" για να προχωράτε παρακάτω μεταξύ των διαλόγων εγκατάστασης.

4. Στο τέλος, όταν εμφανιστεί το παράθυρο διαλόγου που θα λέει "EndNote X is Successfully Installed", πατήστε "Register" ούτως ώστε να καταχωρηθεί το αντίγραφο του EndNote, πατήστε "Finish" ούτως ώστε να κλείσει πρόγραμμα εγκατάστασης, ή πατήστε "Run" για να ξεκινήσετε το EndNote.

.

B.2

Χρησιμοποιώντας το Σύστημα

Εισαγωγή

Σκοπός του κεφαλαίου αυτού είναι να δείξει στο χρήστη πως μπορεί μέσω του εργαλείο EndNote X να αναζητήσει σε βιβλιογραφικές βάσεις δεδομένων στο διαδίκτυο με ευκολία και με τα αποτελέσματα να φτιάξει τις δικές του βάσεις δεδομένων που θα μπορεί να αποθηκεύσει στον ηλεκτρονικό υπολογιστή.

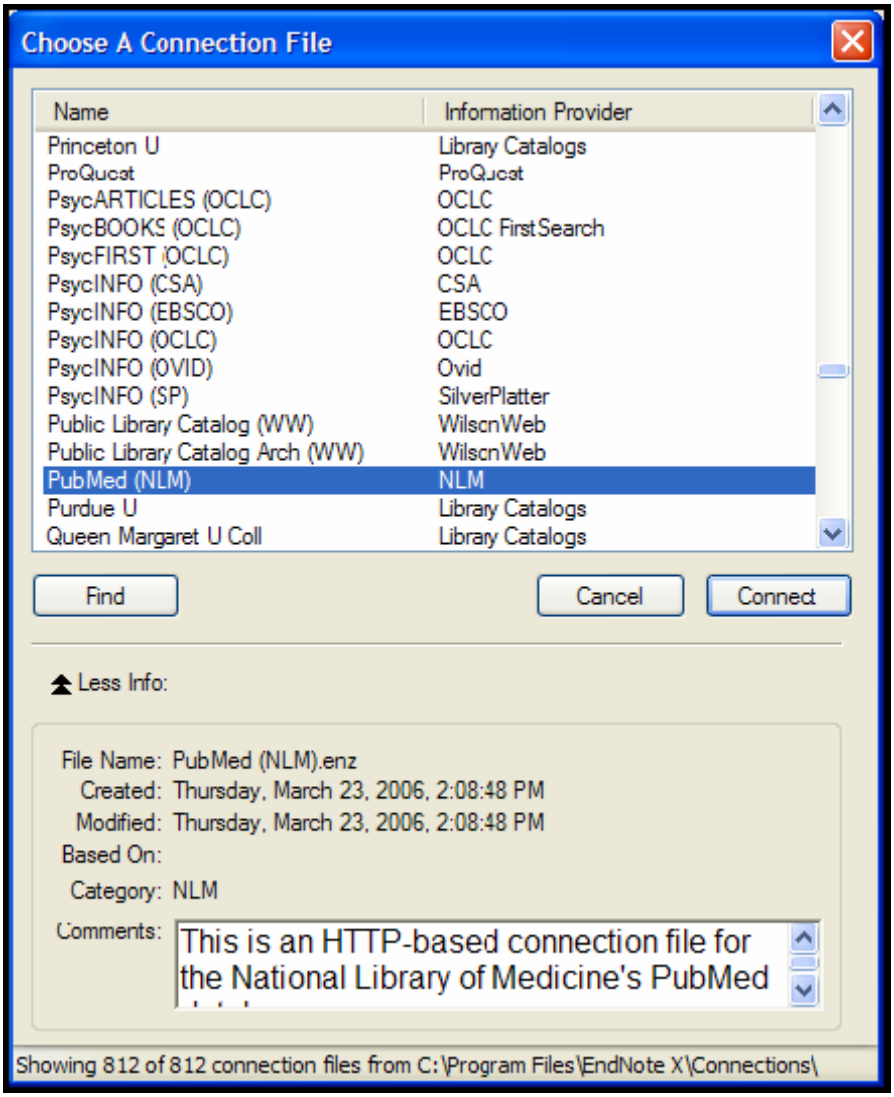
2.1 Σύνδεση με απομακρυσμένη βάση δεδομένων

Το πρώτο βήμα για αναζήτηση σε μια απομακρυσμένη βάση δεδομένων είναι να συνδεθούμε πρώτα με αυτή. Παρακάτω βλέπετε τον τρόπο:

Για τους σκοπούς αυτής της Διπλωματικής Εργασίας ενωθήκαμε με την PubMed.

Έτσι λοιπόν, για να ενωθούμε με τη βάση δεδομένων PubMed:

1. Με το EndNote να είναι σε λειτουργία, πάμε κάτω από το μενού **Tools**, επιλέγουμε το υπό – μενού **Connect**, και στη συνέχεια **Connect**.

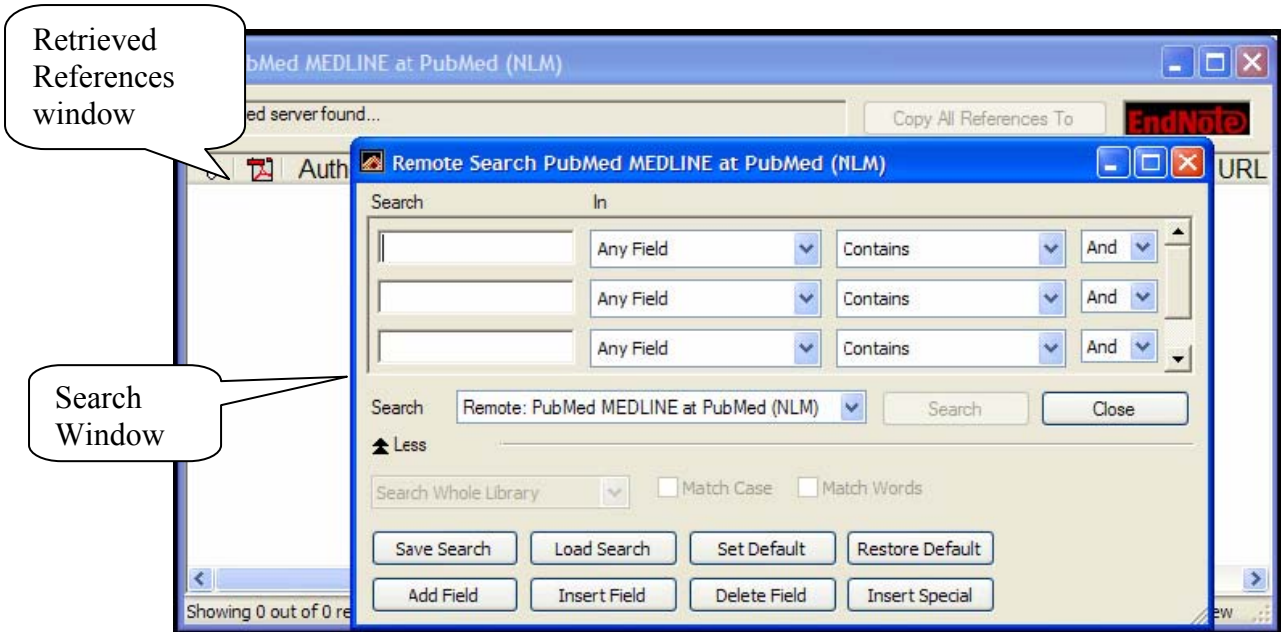


Εικόνα 1: Αυτό το παράθυρο εμφανίζει όλες τις βάσεις με τις οποίες μπορούμε να συνδεθούμε. Μπορείτε να χρησιμοποιήσετε και το κουμπί "Find" για να βρείτε πιο γρήγορα αυτό που ψάχνετε.

2. Επιλέξετε το PubMed αρχείο σύνδεσης και πατήστε "Connect". Επιλέγοντας αυτό το αρχείο σύνδεσης έχετε άμεση πρόσβαση σε αυτή τη βάση δεδομένων.

   Όταν η σύνδεση έχει ολοκληρωθεί, το EndNote εμφανίζει ένα "Retrieved References" παράθυρο για τη βάση δεδομένων PubMed, και τοποθετεί ένα παράθυρο αναζήτησης πάνω από αυτό. Το παράθυρο αναζήτησης το ονομάζει "Remote Search PubMed MEDLINE At PubMed (NLM)."

   Η βάση δεδομένων έχει επιλεχθεί και το EndNote είναι έτοιμο για αναζήτηση.



Εικόνα 3: Το παράθυρο αναζήτησης

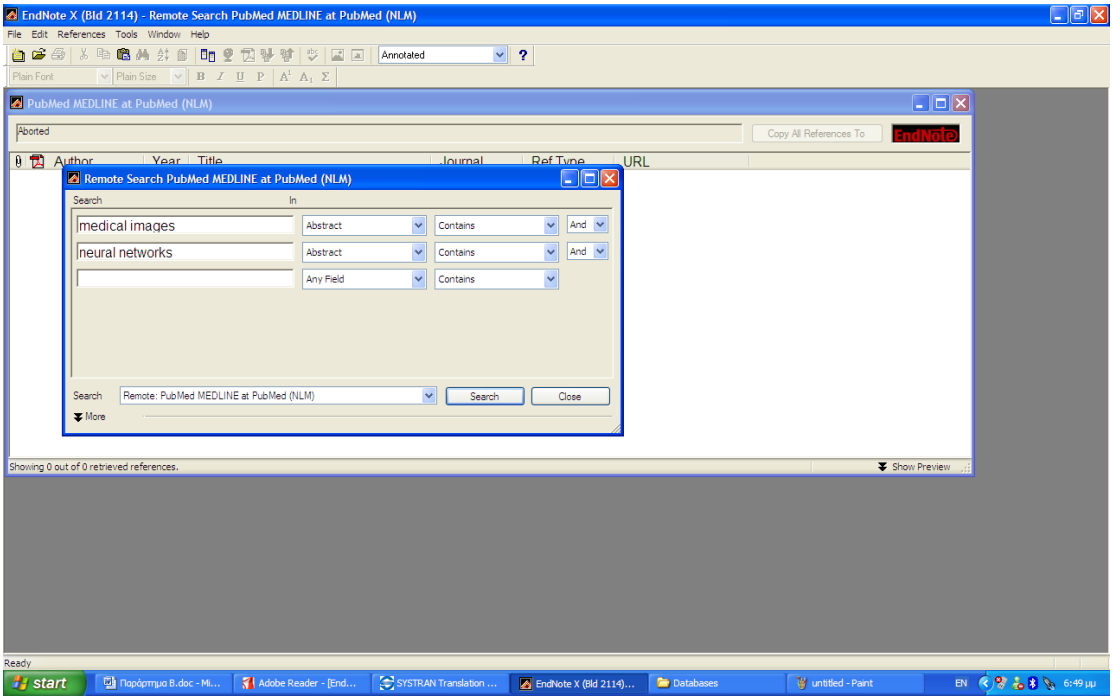
Το επόμενο στάδιο είναι να εισάγετε τον όρο(υς) αναζήτησης για να βρείτε τις αναφορές που επιθυμείτε.

Όπως βλέπετε και από την Εικόνα 3, ο χρήστης μπορεί να κάμει πολλούς συνδυασμούς στην αναζήτηση του (τελεστές AND, OR κ.τ.λ).

2.2 Παράδειγμα δημιουργίας βάσης δεδομένων

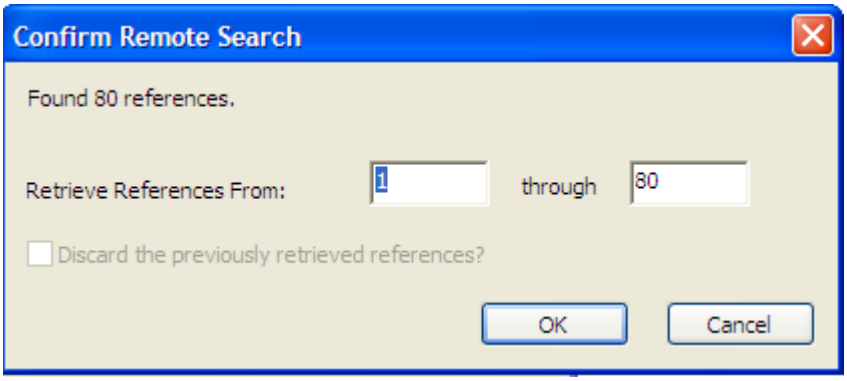
Θέλω να φτιάξω μια βάση δεδομένων για να τη χρησιμοποιήσω σαν είσοδο στο σύστημα που μελετάμε, το σύστημα OntoGen.

Έστω κάμνω αναζήτηση για τους όρους “medical imaging and neural network”



Εικόνα 4: Εισάγω τους όρους medical images and neural networks και θέτω το πεδίο αναζήτησης "Abstract".

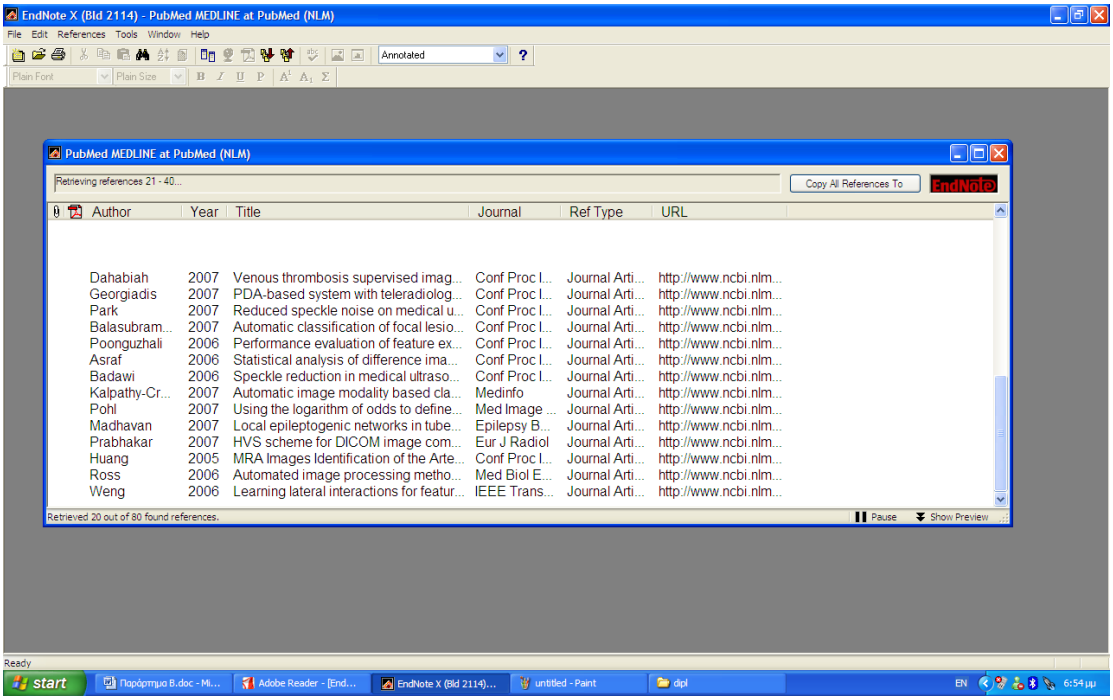
Αφού εισάγω τους όρους για τους οποίους θέλω να γίνει η αναζήτηση, πατάω "Search". Το EndNote στέλνει την απαίτηση μας για αναζήτηση στη συγκεκριμένη βάση και έπειτα με το τέλος της αναζήτησης μας στέλνεται μια περίληψη της αναζήτησης, όπως φαίνεται στην Εικόνα 5:



Εικόνα 5: Ο αριθμός που εμφανίζεται στο παράθυρο κάθε φορά που κάμνουμε μια αναζήτηση μπορεί να διαφέρει για το λόγο ότι η βάση δεδομένων PubMed, καθώς και άλλες βάσεις στο διαδίκτυο ενημερώνονται συνεχώς.

Το παράθυρο εμφανίζει τον αριθμό των αναφορών που βρέθηκαν να ταιριάζουν με τις απαιτήσεις αναζήτησης μας και έπειτα μας δίνεται η επιλογή να τις ανακτήσουμε.

Πατώντας OK οι αναφορές αρχίζουν να ανακτώνται.



Εικόνα 6: Οι αναφορές όπως μεταφέρονται από τη βάση PubMed στο EndNote.

Όταν όλες οι αναφορές ανακτηθούν μπορούμε να τις αντιγράψουμε σε μια νέα βιβλιοθήκη που δημιουργούμε εκείνη τη στιγμή, η οποία θα αποτελέσει μια νέα βάση δεδομένων.

Η βάση δεδομένων θα είναι η είσοδος μας για το σύστημα OntoGen. Μια συλλογή από κείμενα δηλαδή.

Μια διαδικασία η οποία παίρνει ελάχιστο χρόνο και υλοποιείται εύκολα και απλά.