ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Διερεύνηση ιατρικών υποθέσεων για τον καρκίνο του τραχήλου της μήτρας με χρήση προηγμένων τεχνικών εξόρυξης γνώσης σε κείμενα διαδικτυακών βάσεων βιοιατρικής βιβλιογραφίας

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Παναγιώτης Σ. Κατρακάζας

Επιβλέπων : Δημήτριος – Διονύσιος Κουτσούρης

Καθηγητής Ε.Μ.Π

Αθήνα, Δεκέμβριος 2013

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Διερεύνηση ιατρικών υποθέσεων για τον καρκίνο του τραχήλου της μήτρας με χρήση προηγμένων τεχνικών εξόρυξης γνώσης σε κείμενα διαδικτυακών βάσεων βιοιατρικής βιβλιογραφίας

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Παναγιώτης Σ. Κατρακάζας

Επιβλέπων : Δημήτριος – Διονύσιος Κουτσούρης

Καθηγητής Ε.Μ.Π

Εγκρίθηκε από τριμελή εξεταστική επιτροπή την Τρίτη, 7 Ιανουαρίου 2014

……	……	…….
Δ. -Δ. Κουτσούρης	Δ. Φωτιάδης	Γ. Ματσόπουλος
Καθηγητής Ε.Μ.Π	Καθηγητής Παν. Ι.	Επ. Καθηγητής Ε.Μ.Π

Αθήνα, Δεκέμβριος 2013

.....

Παναγιώτης Σ. Κατρακάζας

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π

Πίνακας Περιεχομένων

Ευρετήριο Εικόνων

Ευρετήριο Πινάκων

Περίληψη

Ο στόχος της διπλωματικής εργασίας είναι η διερεύνηση ιατρικών υποθέσεων με προηγμένες τεχνικές εξόρυξης γνώσης από βιοιατρικά κείμενα και η εφαρμογή των τεχνικών αυτών για την αναζήτηση ενός πλαισίου σχέσεων ανάμεσα στον ιό των ανθρωπίνων θηλωμάτων (HPV) και τον καρκίνο των πνευμόνων.

Ο καρκίνος του τραχήλου της μήτρας, αποτελεί την δεύτερη πιο κοινή και την πέμπτη στατιστικά θανατηφόρα μορφή καρκίνου στις γυναίκες[1]. Προσβάλλει περίπου 16 στις 100.000 γυναίκες και προκαλεί θνησιμότητα σε 9 από 100.000 γυναίκες ετησίως[2]. Η ανακάλυψη ότι ο ιός των ανθρωπίνων θηλωμάτων (Human Papillomavirus: HPV) είναι η αποκλειστική αιτία εμφάνισης καρκίνου στο επιθήλιο του τραχήλου της μήτρας, καθιστά τον καρκίνο του τραχήλου της μήτρας μοναδικό στο είδος του.

Οι ραγδαίες τεχνολογικές εξελίξεις στον τομέα της εξόρυξης γνώσης από δεδομένα και κείμενα τα τελευταία χρόνια, κατέστησαν δυνατή τη δημιουργία ενός πλαισίου αναζήτησης και εντοπισμού χρήσιμων πληροφοριών σε βιβλιογραφικές βάσεις δεδομένων, ξεκλειδώνοντας νέες δυνατότητες όσον αφορά την ανακάλυψη γνώσης. Στη βιοιατρική και ειδικότερα όσον αφορά τη διερεύνηση και διατύπωση υποθέσεων, η συμβολή των εργαλείων εξόρυξης κειμένου, είναι παραπάνω από ουσιαστική, καθότι βρίσκουν ολοένα και περισσότερη εφαρμογή με την πάροδο των χρόνων.

Στα πλαίσια της παρούσας εργασίας, παρουσιάσαμε, υπό κάποιες παραδοχές, την εφαρμογή του μοντέλου υποθέσεων του Swanson και σε συνδυασμό με τη χρήση προηγμένων τεχνικών εξόρυξης γνώσης από κείμενα που βρίσκονται σε διαδικτυακές βάσεις βιοιατρικών δεδομένων, μελετήσαμε την πιθανή ύπαρξη σχέσης ανάμεσα στον ιό HPV, που προκαλεί τον καρκίνο του τραχήλου της μήτρας και τον καρκίνο των πνευμόνων. Επικεντρωθήκαμε στην οριοθέτηση του μοντέλου της υπόθεσης και στην αναζήτηση όρων που πιθανώς να συνδέουν τις δυο βιβλιογραφίες. Τα αποτελέσματα στα οποία καταλήξαμε, δείχνουν ότι υπάρχουν πιθανές συσχετίσεις, λόγω κάποιων λέξεων-κλειδιών, οι οποίες χρήζουν περαιτέρω έρευνας.

Λέξεις-Κλειδιά: Εξόρυξη Γνώσης από Κείμενα, Διατύπωση Υποθέσεων, Μοντέλο Υποθέσεων Swanson, Καρκίνος του Τραχήλου της Μήτρας, Ιός Ανθρωπίνων Θηλωμάτων, Καρκίνος των Πνευμόνων

Abstract

The goal of this diploma thesis is to investigate medical hypotheses with advanced text-mining techniques from biomedical texts and the application of these techniques to the search of a framework that indicates possible relations between the Human Papillomavirus (HPV) and lung cancer.

Worldwide, cervical cancer is the second most common cancer and the fifth deadliest one in women. It affects about 16 in 100,000 women per year and kills about 9 in 100,000 per year. The discovery that the Human Papillomavirus (HPV) is the sole cause of cancer in the cervical epithelium, makes cervical cancer unique in its kind.

The rapid technological developments in the field of extracting knowledge from data and texts in recent years, made it possible to create a framework for tracking and searching useful information in bibliographic databases, unlocking new possibilities as far the discovery of knowledge is concerned. In biomedicine, especially from the aspect of investigating and generating hypotheses, the contribution of text mining tools is more than substantial, while they are applied in a more constant basis over the years.

In the context of this work, we presented the implementation of Swanson's Hypothesis Model under some assumptions and combined it with the use of advanced text-mining techniques in texts found in online databases of biomedical data in order to investigate the possible existence of a relationship between the HPV virus and lung cancer. We concentrated on the establishment of the model of this particular case and on the search for keywords that can possibly be found in both bibliographies. Our results indicate that there are potential correlations via certain keywords, which require further investigation.

Key-words: Text-mining, Hypotheses Generation, Swanson's Hypothesis Model, Cervical Cancer, Human Papillomavirus, Lung Cancer

Ευχαριστίες

Η παρούσα διπλωματική εργασία πραγματοποιήθηκε στο Εργαστήριο Βιοιατρικής Τεχνολογίας της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου.

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή, κ. Δημήτριο Κουτσούρη, για την εμπιστοσύνη που έδειξε, αναθέτοντας μου το συγκεκριμένο ενδιαφέρον και επίκαιρο θέμα και για την ευκαιρία που μου έδωσε να ασχοληθώ εμπεριστατωμένα με αυτό. Θα ήθελα επίσης να ευχαριστήσω θερμά τον υποψήφιο Διδάκτορα Χάρη Τσίρμπα για την εξαιρετική συνεργασία που είχαμε όλο αυτό το διάστημα, καθώς και για τον πολύτιμο χρόνο που διέθεσε, προσφέροντας μου βοήθεια και σημαντικές συμβουλές πάνω στην υλοποίηση και διεκπεραίωση της παρούσας εργασίας.

Θα ήθελα επιπλέον να ευχαριστήσω θερμά την υποψήφια Διδάκτορα Βασιλική Πουλή, για τις εποικοδομητικές συμβουλές της και τον χρόνο που διέθεσε καθοδηγώντας με σε αρκετά θέματα της παρούσας διπλωματικής εργασίας, καθώς και την υποψήφια Διδάκτορα Δήμητρα Πουλή και τον ιατρό Σωτήρη Κωνσταντακόπουλο, για την ουσιαστική συμβολή τους όσον αφορά το ιατρικό πλαίσιο της παρούσας εργασίας.

Τέλος, ένα μεγάλο ευχαριστώ σε όλη την οικογένειά μου και στους φίλους μου, για την αμέριστη στήριξη και συμπαράστασή τους καθ΄ όλη την διάρκεια των σπουδών μου και κυρίως στους γονείς μου, που δεν σταμάτησαν να πιστεύουν σε μένα.

Αθήνα, Δεκέμβριος 2013

Κατρακάζας Παναγιώτης

Πρόλογος

Αντικείμενο της παρούσας διπλωματικής εργασίας αποτελεί η μελέτη και εφαρμογή προηγμένων τεχνικών εξόρυξης γνώσης από βιοιατρικά κείμενα πάνω στη διερεύνηση ιατρικών υποθέσεων και συγκεκριμένα, στην αναζήτηση πιθανής βιβλιογραφικής σχέσης ανάμεσα στον ιό των ανθρωπίνων θηλωμάτων (HPV) και στον καρκίνο των πνευμόνων.

Η δομή της εργασίας συνοψίζεται στις εξής ενότητες:

Στο πρώτο κεφάλαιο γίνεται μια εισαγωγή στην εξόρυξη δεδομένων και στην εξόρυξη δεδομένων από κείμενα, όπου παρουσιάζονται οι βασικές αρχές και μεθοδολογίες.

Στο δεύτερο κεφάλαιο γίνεται μια εκτενής αναφορά στην εξόρυξη γνώσης από κείμενα βιοιατρικής βιβλιογραφίας και παρουσιάζονται τα υπάρχοντα εργαλεία, τα οποία διατίθενται για τα διάφορα στάδια των διαδικασιών που περιλαμβάνονται στην εξόρυξη κειμένων.

Στο τρίτο κεφάλαιο περιγράφονται αναλυτικά η σημασία, οι αρχές και το ιστορικό πλαίσιο εξέλιξης της διατύπωσης υποθέσεων στη βιοιατρική και απαριθμώνται τα διάφορα μοντέλα που αναπτύχθηκαν για τον σκοπό αυτό .

Στο τέταρτο κεφάλαιο παρουσιάζεται διεξοδικά η εφαρμογή ενός μοντέλου διερεύνησης γνωστών ιατρικών υποθέσεων που αφορούν στοχευμένα τον καρκίνο του τραχήλου της μήτρας, έπειτα από μια σύντομη επισκόπηση της συγκεκριμένης νόσου και υλοποιείται διερεύνηση για μια νέα υπόθεση που να συνδέει τον ιό Human Papillomavirus με τον καρκίνο των πνευμόνων.

Τέλος, στο πέμπτο κεφάλαιο εξάγονται τα σχετικά συμπεράσματα με βάση όλα τα προαναφερθέντα στοιχεία, παρουσιάζονται οι περιορισμοί της παρούσας εργασίας και προτείνονται οι πιθανές βελτιώσεις για μελλοντική έρευνα.

Κεφάλαιο 1: Εξόρυξη Δεδομένων και Εξόρυξη Δεδομένων από Κείμενα

1.1 Εισαγωγή

Η εξόρυξη δεδομένων (data mining) έχει προσελκύσει τεράστιο ενδιαφέρον τα τελευταία χρόνια στην βιομηχανία της πληροφόρησης αλλά και στην κοινωνία ως σύνολο, λόγω της ευρείας διαθεσιμότητας τεράστιου όγκου δεδομένων και της άμεση ανάγκης που προκύπτει, ώστε να μετατραπούν τα παραπάνω δεδομένα σε χρήσιμη πληροφορία και γνώσεις. Οι πληροφορίες και οι γνώσεις που αποκομίζονται μπορούν να χρησιμοποιηθούν σε ένα ευρύ φάσμα εφαρμογών, το οποίο εκτείνεται από την ανάλυση του marketing, τον εντοπισμό ηλεκτρονικής απάτης και την διατήρηση πελατολογίου έως τον έλεγχο της παραγωγικής διαδικασίας και την εξερεύνηση σε διάφορους τομείς επιστημονικών πεδίων.

Η εξόρυξη δεδομένων μπορεί να θεωρηθεί σαν φυσιολογικό αποτέλεσμα της εξέλιξης της πληροφόρησης. Τα συστήματα διαχείρισης βάσεων δεδομένων έχουν εξελιχθεί σταδιακά όσον αφορά τις παρακάτω λειτουργίες: *συλλογή δεδομένων* και *δημιουργία βάσης δεδομένων*, *διαχείριση δεδομένων* (συμπεριλαμβανομένης της *αποθήκευσης* δεδομένων και της *ανάκτησής* των), *δοσοληψία μεταξύ βάσεων δεδομένων* και *προχωρημένη ανάλυση δεδομένων* (συμπεριλαμβανομένης της *συσσώρευσης τεράστιων όγκων δεδομένων* και της *εξόρυξης δεδομένων*).

Η σταθερή και αλματώδης πρόοδος τις τελευταίες τρεις δεκαετίες, όσον αφορά την τεχνολογία υλικού (hardware) στους υπολογιστές, οδήγησε σε τεράστιες ποσότητες ισχυρών και ταυτόχρονα οικονομικά προσιτών υπολογιστών, καθώς και σε εξοπλισμούς συλλογής δεδομένων και αποθηκευτικών μέσων. Η τεχνολογία αυτή παρέχει ένα μεγάλο άλμα στην βιομηχανία των βάσεων δεδομένων και της πληροφόρησης και κάνει διαθέσιμο έναν τεράστιο αριθμό από βάσεις δεδομένων και ‘’αποθηκών’’ πληροφορίας για συνδιαλλαγή, ανάκτηση πληροφοριών και ανάλυση δεδομένων.

Η αφθονία των δεδομένων, σε συνδυασμό με την ανάγκη για ισχυρά εργαλεία ανάλυσης δεδομένων, περιγράφεται ως μια κατάσταση *πλούσια σε δεδομένα, αλλά φτωχή σε πληροφορία*. Ο ολοένα αυξανόμενος όγκος δεδομένων, που συλλέγεται και αποθηκεύεται σε πολυάριθμες, τεράστιες αποθήκες δεδομένων, έχει ξεπεράσει κατά πολύ την ανθρώπινη ικανότητα κατανόησής των, χωρίς την βοήθεια ισχυρών εργαλείων. Σαν αποτέλεσμα, δεδομένα που συσσωρεύονται σε τέτοιες αποθήκες καταλήγουν να μετατρέπονται σε *τάφους δεδομένων*, δηλαδή δεδομένα τα οποία προσπελαύνονται σπανίως. Κατά συνέπεια αποφάσεις που χρήζουν σημασίας, δεν λαμβάνονται βάσει της χρήσιμης πληροφορίας που περιέχεται σε δεδομένα, τα οποία υπάρχουν σε αποθήκες δεδομένων, αλλά βάσει του ενστίκτου του χρήστη, ο οποίος δεν έχει τα απαραίτητα εργαλεία ώστε να εξάγει την χρήσιμη γνώση που περιέχεται σε τεράστιους όγκους δεδομένων. Ακόμα όμως και τα πιο εξειδικευμένα τεχνολογικά συστήματα βασίζονται στους χρήστες ή στους ειδικευμένους τεχνικούς, οι οποίοι εισάγουν *χειροκίνητα* τις χρήσιμες και απαραίτητες πληροφορίες στις βάσεις δεδομένων τους, μια διαδικασία η οποία είναι επιρρεπής σε λάθη και υποκειμενικότητες, ενώ ταυτόχρονα είναι χρονοβόρα και οικονομικά ασύμφορη. Τα εργαλεία εξόρυξης δεδομένων παρέχουν ανάλυση των δεδομένων και μπορούν να αποκαλύψουν πιθανόν σημαντικές αλληλουχίες – πρότυπα δεδομένων, συμβάλλοντας κατά μεγάλο βαθμό σε επιχειρηματικές στρατηγικές, επιστημονικές και ιατρικές έρευνες και σε βέλτιστες βάσεις γνώσεων. Το κενό που υπάρχει ανάμεσα στα δεδομένα και στην γνώση

απαιτεί μια συστηματική ανάπτυξη των εργαλείων εξόρυξης δεδομένων που θα μετατρέπουν τους *τάφους δεδομένων* σε *ψήγματα γνώσεως*.

1.2 Ορισμός Εξόρυξης Δεδομένων (Data Mining)

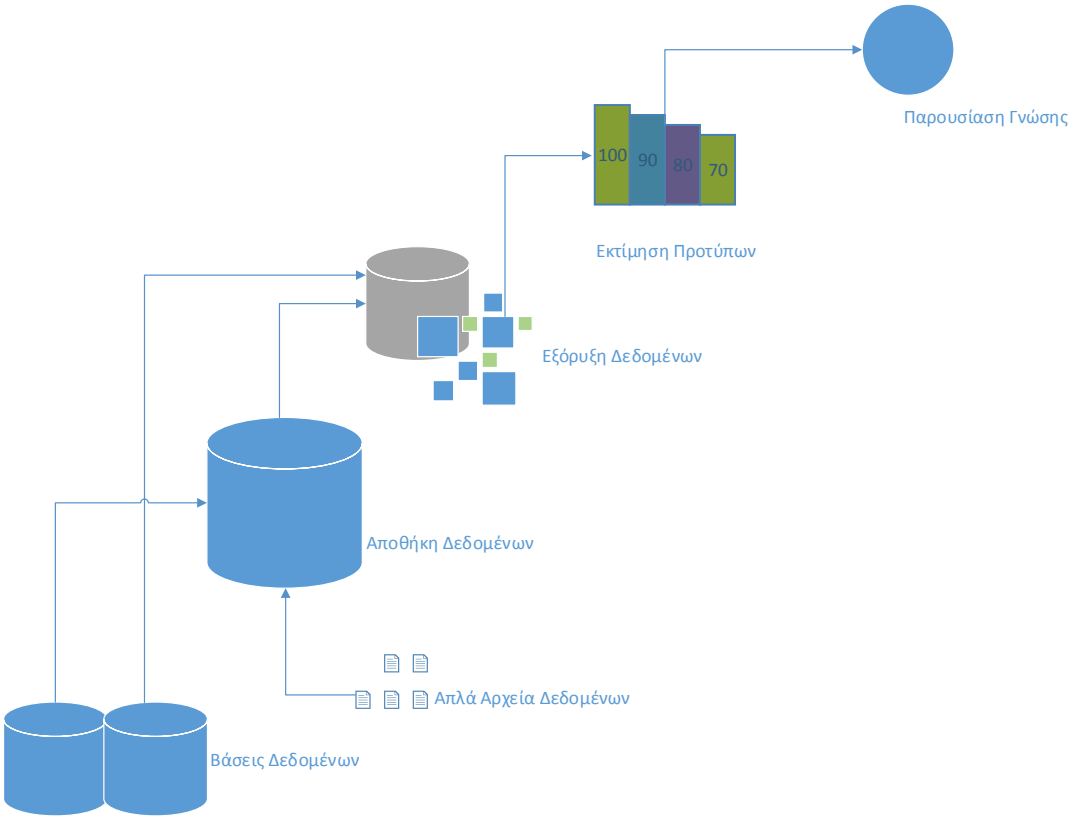
Η εξόρυξη δεδομένων αναφέρεται στην εξαγωγή ή «εξόρυξη» γνώσης από μεγάλο όγκο δεδομένων. Ο όρος αυτός είναι στην ουσία εσφαλμένος, διότι στην πραγματικότητα γινόταν εξόρυξη χρυσού και όχι εξόρυξη πέτρας ή άμμου στα χρυσωρυχεία. Οπότε ουσιαστικά έπρεπε να αναφέρεται ως *εξόρυξη γνώσης*, κάτι όμως το οποίο δεν αντικατοπτρίζει την έμφαση που δίνεται στο μεγάλο όγκο δεδομένων από τα οποία γίνεται η εξόρυξη. Κατά συνέπεια, κατέληξε να χρησιμοποιείται ο παραπάνω όρος πιο συχνά σε σχέση με άλλους παραπλήσιους προς αυτόν, όπως *εξαγωγή πληροφορίας* και *ανάλυση προτύπων δεδομένων*.

Συνήθως ο όρος εξόρυξη δεδομένων θεωρείται συνώνυμο ενός άλλου δημοφιλούς όρου, της Ανακάλυψης Γνώσης από Βάσεις Δεδομένων, ή αλλιώς KDD (Knowledge Discovery from Data), ενώ άλλες φορές θεωρείται σαν ένα απαραίτητο βήμα στην διαδικασία για την ανακάλυψη της χρήσιμης πληροφορίας. Η Ανακάλυψη Γνώσης σαν διαδικασία απεικονίζεται στην **Εικόνα 1** και αποτελεί μια επαναληπτική ακολουθία των παρακάτω σταδίων:

1. **Εκκαθάριση Δεδομένων** (εξάλειψη θορύβου και άσχετων δεδομένων)
2. **Ενσωμάτωση Δεδομένων** (πιθανότητα συνδυασμού πολλαπλών πηγών δεδομένων)
3. **Επιλογή Δεδομένων** (ανάκτηση από τις βάσεις εκείνων των δεδομένων που σχετίζονται με το θέμα ανάλυσης)
4. **Μετατροπή Δεδομένων** (μετασχηματισμός ή συνένωση των δεδομένων σε μορφές κατάλληλες για εξόρυξη, εκτελώντας αθροιστικές ή περιληπτικές λειτουργίες για παράδειγμα)
5. **Εξόρυξη δεδομένων** (μια απαραίτητη διαδικασία όπου ευφυείς μέθοδοι εφαρμόζονται ώστε να εξαγάγουν πρότυπα δεδομένων)
6. **Εκτίμηση Προτύπων** (αναγνώριση εκείνων των προτύπων που αναπαριστούν γνώση βασισμένη σε κριτήρια ενδιαφέροντος)
7. **Παρουσίαση Γνώσης** (τεχνικές απεικόνισης και αναπαράστασης γνώσης χρησιμοποιούνται ώστε να παρουσιάσουν την εξορυγμένη γνώση στον χρήστη)

Τα στάδια 1 έως 4 αποτελούν διαφορετικές μορφές προ-επεξεργασίας δεδομένων, όπου τα δεδομένα προετοιμάζονται για εξόρυξη. Η εξόρυξη δεδομένων μπορεί να αλληλεπιδρά με τον χρήστη ή με μια βάση γνώσης. Τα πρότυπα που έχουν ενδιαφέρον παρουσιάζονται στον χρήστη και μπορεί να αποθηκευτούν σαν νέα γνώση στην βάση γνώσης. Σύμφωνα με την παραπάνω προσέγγιση, η εξόρυξη δεδομένων είναι ένα μόνο στάδιο στην όλη διαδικασία, εντούτοις σημαντικό, διότι αποκαλύπτει κρυμμένα πρότυπα για αξιολόγηση.

Ακολουθώντας μια διαφορετική προσέγγιση, η εξόρυξη δεδομένων μπορεί να θεωρηθεί ανεξάρτητα ως η διαδικασία ανακάλυψης ενδιαφέρουσας γνώσης από μεγάλους όγκους δεδομένων αποθηκευμένων σε βάσεις δεδομένων, αποθήκες δεδομένων ή άλλες αποθήκες πληροφοριών. Σύμφωνα με αυτή την προσέγγιση, η αρχιτεκτονική ενός τυπικού συστήματος εξόρυξης δεδομένων, μπορεί να περιλαμβάνει τα παρακάτω κύρια στοιχεία (**Εικόνα 2**):



Εικόνα 1 Η εξόρυξη δεδομένων ως στάδιο στη διαδικασία ανακάλυψης γνώσης

- **Βάση Δεδομένων, Αποθήκη Δεδομένων, Παγκόσμιος Ιστός (World Wide Web) ή κάποια άλλη αποθήκη πληροφοριών** : Αποτελείται από μεμονωμένες ή ένα σύνολο από βάσεις δεδομένων, αποθήκες πληροφοριών, λογιστικά φύλλα ή κάποιο άλλο είδος αποθηκών πληροφορίας. Τεχνικές εκκαθάρισης και ενσωμάτωσης δεδομένων μπορούν να εφαρμοστούν στα δεδομένα.

- **Διακομιστής Βάσης Δεδομένων ή Αποθήκης Δεδομένων** : Ο διακομιστής είναι υπεύθυνος για την προσκόμιση των σχετικών δεδομένων, σύμφωνα με το αίτημα εξόρυξης δεδομένων του χρήστη.

- **Βάση Γνώσης** : Είναι η κύρια γνώση που χρησιμοποιείται για να καθοδηγήσει την αναζήτηση ή την εκτίμηση του ενδιαφέροντος για τα πρότυπα που προέκυψαν. Η γνώση αυτή μπορεί να περιλαμβάνει **ιεραρχίες εννοιών**, που χρησιμοποιούνται για να αποδώσουν χαρακτηριστικά ή ιδιότητες χαρακτηριστικών σε διαφορετικά αφαιρετικά επίπεδα. Μπορεί επίσης να χρησιμοποιηθεί γνώση όπως πεποιθήσεις χρηστών, που μπορούν να χρησιμοποιηθούν για την εκτίμηση του ενδιαφέροντος ενός προτύπου βάσει του απροσδόκητου του.

- **Μηχανή εξόρυξης δεδομένων** : Απαραίτητο στοιχείο στο σύστημα εξόρυξης δεδομένων και ιδανικά αποτελείται από ένα σύνολο λειτουργικών ενοτήτων με έργο τον χαρακτηρισμό, την ανάλυση για σύνδεση και συσχέτιση δεδομένων, την

ταξινόμηση, πρόβλεψη και ανάλυση συστάδων δεδομένων, την εξέλιξη και ανάπτυξη απομακρυσμένων δεδομένων.

- **Μονάδα Εκτίμησης Προτύπων** : Το τμήμα αυτό τυπικά αποδίδει κριτήρια ενδιαφέροντος και αλληλεπιδρά με την μηχανή εξόρυξης δεδομένων, ώστε να επικεντρωθεί η αναζήτηση προς τα πρότυπα που ενδιαφέρουν. Υπάρχει δυνατότητα χρήσης «κατωφλιών ενδιαφέροντος» (interestingness thresholds) ώστε να φιλτραριστούν τα ανακαλυφθέντα πρότυπα. Εναλλακτικά, η μονάδα εκτίμησης προτύπων μπορεί να ενσωματωθεί με την μηχανή εξόρυξης δεδομένων, ανάλογα με την υλοποίηση της μεθόδου εξόρυξης δεδομένων που χρησιμοποιείται. Για αποτελεσματικότερη εξόρυξη δεδομένων συνίσταται η εμβάθυνση όσο το δυνατόν περισσότερο, της αποτίμησης του ενδιαφέροντος σε συγκεκριμένα πρότυπα στην διαδικασία της εξόρυξης έτσι ώστε να περιοριστεί η αναζήτηση για τα πρότυπα που μας ενδιαφέρουν.

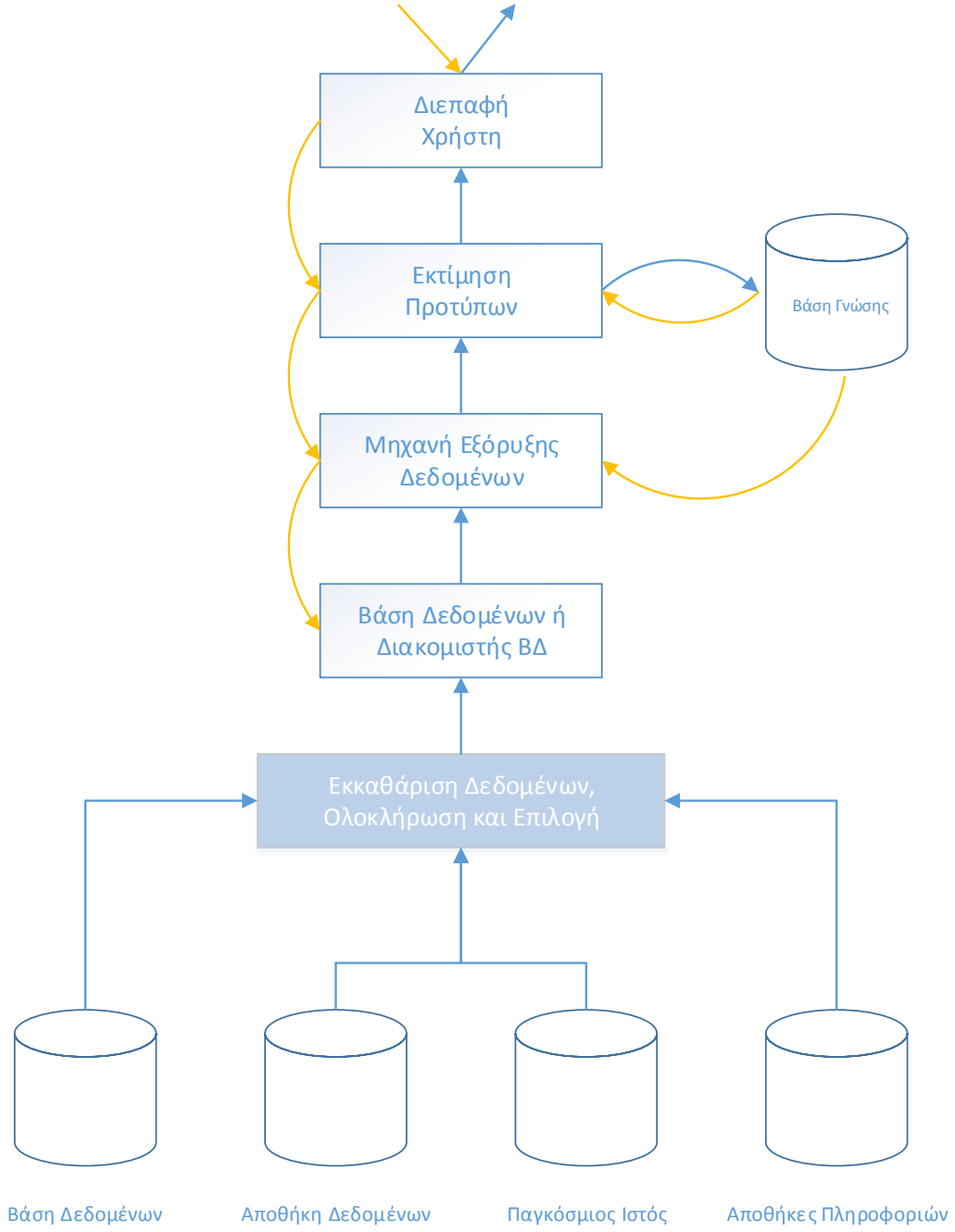
- **Διεπαφή Χρήστη :** Το τμήμα αυτό βοηθάει στην επικοινωνία μεταξύ χρηστών και του συστήματος εξόρυξης δεδομένων. Επιτρέπει στον χρήστη να αλληλεπιδρά με το σύστημα αναθέτοντας ένα συγκεκριμένο ερώτημα ή θέμα για εξόρυξη δεδομένων, παρέχοντας πληροφορίες που βοηθάνε την εστίαση της αναζήτησης και εκτελώντας εξερευνητικές εξορύξεις δεδομένων βασισμένες στα ενδιάμεσα αποτελέσματα. Επιπλέον, επιτρέπει στον χρήστη να έχει πρόσβαση σε βάσεις δεδομένων ή αποθήκες δεδομένων, να αξιολογήσει εξορυγμένα πρότυπα και να οπτικοποιήσει τα πρότυπα σε διαφορετικές μορφές.

Απ' την άλλη πλευρά, παρόλο που στην αγορά μπορεί να κυκλοφορούν πολλά συστήματα εξόρυξης δεδομένων, δεν εκτελούν όλα πραγματική εξόρυξη δεδομένων. Ένα σύστημα ανάλυσης δεδομένων που δεν χειρίζεται μεγάλες ποσότητες δεδομένων, είναι καλύτερα να χαρακτηρίζεται ως σύστημα εκμάθησης μηχανής, ένα εργαλείο στατιστικής ανάλυσης ή ένα πειραματικό πρωτότυπο σύστημα. Ένα σύστημα που κάνει μόνο ανάκτηση δεδομένων ή πληροφοριών, συμπεριλαμβανομένης της εύρεσης συνολικών τιμών, ή που εκτελεί επαγωγική απάντηση ερωτημάτων σε μεγάλες βάσεις δεδομένων, μπορεί καταλληλότερα να χαρακτηριστεί ως σύστημα βάσης δεδομένων ή σύστημα ανάκτησης πληροφοριών ή ένα επαγωγικό σύστημα δεδομένων.

1.3 Εξόρυξη Δεδομένων από Κείμενα (Text data mining)

Μια ουσιαστική ποσότητα των διαθέσιμων πληροφοριών είναι αποθηκευμένες σε βάσεις (δεδομένων) κειμένων, οι οποίες αποτελούνται από μεγάλες συλλογές εγγράφων από διάφορες πηγές, όπως για παράδειγμα άρθρα από εφημερίδες, ερευνητικές δημοσιεύσεις, βιβλία, ψηφιακές βιβλιοθήκες, ηλεκτρονικά μηνύματα και ιστοσελίδες. Οι βάσεις κειμένων διευρύνονται όλο και περισσότερο λόγω της ολοένα αυξανόμενης ποσότητας πληροφοριών που γίνονται διαθέσιμες σε ηλεκτρονική μορφή, όπως ηλεκτρονικές δημοσιεύσεις, διαφόρων ειδών ηλεκτρονικά έγγραφα, ηλεκτρονική αλληλογραφία, αλλά και ο ίδιος ο Παγκόσμιος Ιστός, που μπορεί να θεωρηθεί σαν μια τεράστια, αλληλοσυνδεόμενη, δυναμική βάση κειμένου. Στις μέρες μας, οι περισσότερες πληροφορίες σε κυβερνητικά και

επιστημονικά ιδρύματα, βιομηχανίες, επιχειρήσεις και άλλους οργανισμούς αποθηκεύονται ηλεκτρονικά, στην μορφή βάσεων κειμένων.



Εικόνα 2 Τυπικό Σύστημα Εξόρυξης Δεδομένων

Τα δεδομένα που αποθηκεύονται στις περισσότερες βάσεις κειμένων, είναι ημι-δομημένα, με την έννοια ότι δεν είναι ούτε αδόμητα, αλλά ούτε και τελείως δομημένα. Για παράδειγμα, ένα κείμενο μπορεί να περιέχει ορισμένα δομημένα πεδία, όπως για παράδειγμα τίτλο, συγγραφέα, ημερομηνία έκδοσης, κατηγορία κ.ο.κ. αλλά μπορεί να περιέχει και αδόμητες

μορφές κειμένου, όπως τα περιεχόμενα και την περίληψη. Πρόσφατα, έχουν γίνει πολλές μελέτες που αφορούν την μοντελοποίηση και την βελτίωση των ημι-δομημένων δεδομένων και επιπλέον, έχουν αναπτυχθεί τεχνικές ανάκτησης δεδομένων όπως μέθοδοι τοποθέτησης δεικτών σε κείμενα, ώστε να χειριστούν αδόμητα κείμενα.

Οι παραδοσιακές τεχνικές ανάκτησης πληροφοριών, έχουν γίνει πλέον ανεπαρκείς για τις υπερβολικές ποσότητες κειμένων που υπάρχουν σήμερα. Τυπικά, μόνο ένα μικρό ποσοστό από τα πολλά διαθέσιμα κείμενα θα είναι σχετικό για ένα δεδομένο χρήστη. Χωρίς γνώση του περιεχομένου των εγγράφων, είναι δύσκολο να σχηματιστούν αποτελεσματικά ερωτήματα για ανάλυση και εξαγωγή της χρήσιμης πληροφορίας από τα κείμενα. Οι χρήστες χρειάζονται εργαλεία που να συγκρίνουν τα διάφορα κείμενα, να βαθμολογούν την σημαντικότητα και σχετικότητα τους ή να βρίσκουν πρότυπα και ομοιότητες ανάμεσα σε πολλαπλά κείμενα. Κατ' αυτήν την έννοια, η εξόρυξη δεδομένων από κείμενα έχει γίνει ιδιαίτερα δημοφιλής και απαραίτητη στην εξόρυξη δεδομένων.

1.3.1 Ανάλυση Δεδομένων Κειμένων και Ανάκτηση Πληροφοριών

Η Ανάκτηση Πληροφοριών (Information Retrieval, IR) είναι ένα πεδίο που αναπτύσσεται παράλληλα με τα συστήματα βάσεων δεδομένων. Σε αντίθεση με το πεδίο των συστημάτων βάσεων δεδομένων, που έχει επικεντρωθεί στην επεξεργασία ερωτημάτων και συνδιαλλαγών δομημένων δεδομένων, η ανάκτηση πληροφοριών ασχολείται με την οργάνωση και ανάκτηση των πληροφοριών από έναν μεγάλο αριθμό γραπτών κειμένων. Από τη στιγμή που τα δυο αυτά πεδία, χειρίζονται διαφορετικού τύπου δεδομένα, κάποια προβλήματα των συστημάτων βάσεων δεδομένων, όπως για παράδειγμα ο έλεγχος του συγχρονισμού των δεδομένων, η ανάκτηση, η διαχείριση των συνδιαλλαγών και η ενημέρωση, δεν υπάρχουν στα συστήματα ανάκτησης πληροφοριών. Αντίστοιχα, υπάρχουν προβλήματα των συστημάτων ανάκτησης πληροφοριών, όπως για παράδειγμα, αδόμητα κείμενα, παραπλήσια αναζήτηση βασισμένη σε λέξεις-κλειδιά καθώς και η έννοια της σχετικότητας, που δεν συναντώνται στα συστήματα βάσεων δεδομένων.

Λόγω της αφθονίας γραπτών πληροφοριών, η ανάκτηση πληροφοριών ανταποκρίνεται σε πολλές εφαρμογές, όπως τα συστήματα που χειρίζονται τους online καταλόγους μιας βιβλιοθήκης, αυτά που χειρίζονται online έγγραφα και οι πιο πρόσφατες, δικτυακές μηχανές αναζήτησης (web search engines).

Ένα τυπικό πρόβλημα ανάκτησης πληροφοριών είναι να εντοπιστούν σχετικά έγγραφα σε μια συλλογή εγγράφων βασισμένα στο ερώτημα ενός χρήστη, που είναι συνήθως κάποιες λέξεις – κλειδιά που περιγράφουν μια ανάγκη για συγκεκριμένες πληροφορίες ή ένα σχετικό κείμενο. Σε ένα τέτοιο πρόβλημα αναζήτησης, ο χρήστης παίρνει την πρωτοβουλία να «τραβήξει» τις σχετικές πληροφορίες από τη συλλογή, κάτι που είναι χρήσιμο όταν ο χρήστης έχει ανάγκη από πληροφορίες για κάποια περίσταση βραχυπρόθεσμα, όπως για παράδειγμα την αγορά ενός μεταχειρισμένου αυτοκινήτου. Όταν όμως έχει ανάγκη από πληροφορίες μακροπρόθεσμα, όπως λόγου χάρη για μια επιστημονική έρευνα, ένα σύστημα ανάκτησης πληροφοριών μπορεί να πάρει την πρωτοβουλία να «προωθήσει» οποιαδήποτε νέα πληροφορία στον χρήστη, εφόσον αυτή κριθεί σχετική με τις πληροφοριακές ανάγκες του. Αυτή η διαδικασία πληροφοριακής πρόσβασης καλείται *φιλτράρισμα πληροφοριών* και τα συστήματα που την εκτελούν, *συστήματα φιλτραρίσματος* ή *συμβουλευτικά συστήματα*.

1.3.2 Βασικά Κριτήρια για Ανάκτηση Κειμένων: Ακρίβεια και Ανάκληση

Ας υποθέσουμε ότι ένα σύστημα ανάκτησης κειμένων, έχει επιστρέψει έναν αριθμό εγγράφων, βάσει ενός ερωτήματος που του έχουμε θέσει. Το ερώτημα που τίθεται είναι πως μπορούμε να εκτιμήσουμε την ακρίβεια και τη σωστή λειτουργία του συστήματος. Θεωρούμε ότι το σύνολο των εγγράφων, τα οποία είναι σχετικά με το ερώτημα που θέσαμε, ορίζεται ως ***{Σχετικά}*** και το σύνολο των κειμένων που ανακτήθηκαν ως ***{Ανακτημένα}***. Το σύνολο των εγγράφων που είναι ταυτόχρονα ανακτημένα και σχετικά με το ερώτημα, ορίζεται ως η τομή των δυο παραπάνω συνόλων, σύμφωνα με το διάγραμμα Venn που εμφανίζεται στην **Εικόνα 3**.

Εικόνα 3 Κατηγοριοποίηση Εγγράφων στην Ανάκτηση Κειμένων

Υπάρχουν δυο βασικά κριτήρια για την εκτίμηση της ποιότητας της ανάκτησης κειμένων:

- **Ακρίβεια** : Εκφράζει το ποσοστό των εγγράφων που είναι σχετικά με το ερώτημα (ουσιαστικά οι «σωστές» απαντήσεις στο ερώτημα), από το σύνολο των ανακτημένων εγγράφων. Ορίζεται ως:

$$ακρίβεια = \frac{|\{Σχετικά\} \cap \{Ανακτημένα\}|}{|\{Ανακτημένα\}|}$$

- **Ανάκληση** : Εκφράζει το ποσοστό των εγγράφων που ανακτήθηκαν από το σύνολο των σχετικών με το ερώτημα εγγράφων. Ορίζεται ως:

$$ανάκληση = \frac{|\{Σχετικά\} \cap \{Ανακτημένα\}|}{|\{Σχετικά\}|}$$

Ένα σύστημα ανάκτησης πληροφοριών πολλές φορές χρειάζεται να βρει μια μέση οδό όσον αφορά την χρήση ενός κριτηρίου εις βάρος του άλλου και το αντίθετο. Μια συχνά χρησιμοποιούμενη μέση λύση αποτελεί το ***F_score***, που ορίζεται ως ο αρμονικός μέσος της ακρίβειας και της ανάκλησης:

$$F_{score} = \frac{ακρίβεια \times ανάκληση}{(ακρίβεια + ανάκληση)/2}$$

Ο παραπάνω όρος «αποθαρρύνει» ένα σύστημα να χρησιμοποιήσει ένα κριτήριο δραστικά εις βάρος του άλλου.

Τα παραπάνω κριτήρια (***ακρίβεια***, ***ανάκληση*** και ***F_score***) είναι τα βασικά κριτήρια στην ανάκτηση ενός συνόλου εγγράφων. Δεν είναι όμως άμεσα χρήσιμα για σύγκριση δυο ταξινομημένων λιστών εγγράφων, διότι δεν είναι «ευαίσθητα» ως προς την εσωτερική ταξινόμηση των εγγράφων σε ένα ανακτημένο σύνολο. Για να υπολογίσουμε την ποιότητα μιας ταξινομημένης λίστας εγγράφων συνήθως υπολογίζεται η μέση τιμή της ακρίβειας από όλες τις ταξινομήσεις, όπου ένα νέο, σχετικό έγγραφο έχει προστεθεί. Επιπλέον είναι σύνηθες να σχεδιάζεται ένα γράφημα της ακρίβειας συναρτήσει των πολλών διαφορετικών επιπέδων της ανάκλησης. Όσο υψηλότερη είναι η καμπύλη, τόσο καλύτερης ποιότητας είναι το σύστημα ανάκτησης πληροφοριών.

1.3.3 Μέθοδοι Ανάκτησης Εγγράφων

Οι μέθοδοι ανάκτησης εγγράφων γενικά χωρίζονται σε δυο κατηγορίες, ανάλογα με τον τρόπο αντιμετώπισης του προβλήματος ανάκτησης, είτε ως ένα πρόβλημα ***επιλογής εγγράφων*** είτε ως ένα πρόβλημα ***ταξινόμησης εγγράφων***.

Στις μεθόδους **επιλογής εγγράφων**, το ερώτημα θεωρείται ως ένα σύνολο καθορισμένων περιορισμών για την επιλογή των σχετικών εγγράφων. Μια τυπική μέθοδος αυτής της κατηγορίας είναι το **μοντέλο ανάκτησης Boolean**, στο οποίο ένα έγγραφο αντιπροσωπεύεται από ένα σύνολο λέξεων-κλειδιών και ο χρήστης παρέχει μια λογική έκφραση τύπου Boolean, όπως για παράδειγμα «αυτοκίνητο *ΚΑΙ* συνεργεία», «τσάι *Η* καφές» κτλ. Το σύστημα θα δεχτεί αυτή την έκφραση και θα επιστρέψει τα κείμενα εκείνα που καλύπτουν την αντίστοιχη έκφραση. Εξαιτίας της δυσκολίας που υπάρχει στο να περιγραφούν με ακρίβεια οι πληροφορίες που χρειάζεται ένας χρήστης μέσω των Boolean εκφράσεων, η μέθοδος αυτή λειτουργεί καλά μόνο όταν ο χρήστης έχει επαρκή γνώση της συλλογής εγγράφων που διαθέτει και μπορεί να σχηματίσει καλώς ορισμένα ερωτήματα πάνω σ' αυτά.

Στις **μεθόδους ταξινόμησης**, το ερώτημα χρησιμοποιείται ώστε να ταξινομήσει όλα τα έγγραφα με σειρά σχετικότητας. Για τους μέσους χρήστες και για διερευνητικά ερωτήματα, οι μέθοδοι ταξινόμησης προτιμώνται έναντι των μεθόδων επιλογής εγγράφων. Τα περισσότερα σύγχρονα συστήματα ανάκτησης δεδομένων παρουσιάζουν μια ταξινομημένη λίστα των εγγράφων που απαντούν στο ερώτημα (λέξη-κλειδί) ενός χρήστη. Υπάρχουν αρκετές διαφορετικές μέθοδοι ταξινόμησης που βασίζονται σε ένα μεγάλο εύρος μαθηματικών υποδομών, συμπεριλαμβανομένων της άλγεβρας, της λογικής, της στατιστικής και των πιθανοτήτων. Η κοινή λογική πίσω από όλες αυτές τις μεθόδους είναι να «ταιριάξουν» οι λέξεις-κλειδιά ενός ερωτήματος με αυτές σε ένα έγγραφο και να ταξινομηθεί

το έγγραφο ανάλογα με το πόσο καλά απαντάει το ερώτημα. Ο σκοπός είναι να προσεγγίσει τον βαθμό σχετικότητας ενός εγγράφου βάσει ενός αποτελέσματος, υπολογισμένο με πληροφορίες όπως η συχνότητα των λέξεων μέσα στο έγγραφο. Εντούτοις, είναι εγγενώς δύσκολο να υπάρχει ένα ακριβές μέτρο του βαθμού σχετικότητας ανάμεσα σε ένα σύνολο λέξεων-κλειδιών (για παράδειγμα, είναι δύσκολο να προσδιορισθεί η διαφορά ανάμεσα στην *εξόρυξη δεδομένων* και στην *ανάλυση δεδομένων*). Γι' αυτό το λόγο είναι απαραίτητη μια περιεκτική, εμπειρική εκτίμηση στην επικύρωση οποιασδήποτε μεθόδου ανάκτησης δεδομένων.

1.3.3.1 Μοντέλο Διανυσματικού Χώρου (Vector-Space Model)

Η βασική ιδέα του **Μοντέλου Διανυσματικού Χώρου** είναι η ακόλουθη: το έγγραφο και το ερώτημα αναπαριστώνται ως διανύσματα σε έναν πολυδιάστατο χώρο που ανταποκρίνεται σε όλες τις λέξεις-κλειδιά και χρησιμοποιούνται κριτήρια ομοιότητας, ώστε να υπολογιστεί η ομοιότητα μεταξύ του διανύσματος-ερώτημα και του διανύσματος-έγγραφο. Οι τιμές ομοιότητας μπορούν μετά να χρησιμοποιηθούν για την ταξινόμηση εγγράφων.

Το πρώτο βήμα στα περισσότερα συστήματα ανάκτησης δεδομένων είναι η αναγνώριση λέξεων-κλειδιών για την αντιπροσώπευση εγγράφων, ένα προ-επεξεργαστικό στάδιο που ονομάζεται **κατακερματισμός** (**tokenization**). Για την αποφυγή δεικτοδότησης άχρηστων λέξεων, τα συστήματα ανάκτησης δεδομένων συνδέουν τα κείμενα με μια **stop list**, η οποία αποτελείται από ένα σύνολο λέξεων (οι οποίες αποκαλούνται **stop words**), οι οποίες εμφανίζονται τόσο συχνά μέσα σε ένα κείμενο ώστε να χάνουν την πληροφοριακή τους χρησιμότητα (π.χ. a, are, as, the, for, it, with, have, of, to, will κ.ο.κ.). Η **stop list** μπορεί να διαφέρει ανά σύνολο κειμένων, για παράδειγμα, οι λέξεις «συστήματα δεδομένων» μπορεί να είναι μια σημαντική λέξη-κλειδί σε εφημερίδες, αλλά δεν είναι σημαντική για ερευνητικές δημοσιεύσεις που αφορούν τα συστήματα δεδομένων.

Σύνολα διαφορετικών λέξεων μπορεί να μοιράζονται τον ίδιο λεκτικό κορμό (**word stem**). Ένα σύστημα ανάκτησης κειμένων χρειάζεται να μπορεί να αναγνωρίζει ομάδες λέξεων που έχουν μικρές συντακτικές διαφορές μεταξύ τους και να συλλέγει μόνο τον κοινό λεκτικό κορμό ανά ομάδα. Για παράδειγμα, οι λέξεις *drug*, *drugged* και *drugs* έχουν ως κοινό λεκτικό κορμό το *drug*, οπότε μπορούν να προσμετρηθούν σαν διαφορετικές εμφανίσεις της ίδιας λέξης.

Μαθηματικοποιώντας τα παραπάνω, έστω ότι αρχικά έχουμε ένα σύνολο εγγράφων $\boldsymbol{d}$ και ένα σύνολο όρων $\boldsymbol{t}$. Μπορούμε να μοντελοποιήσουμε το κάθε έγγραφο ως ένα διάνυσμα $\underline{\boldsymbol{v}}$ στον t-διάστατο χώρο $\mathbb{R}^{\boldsymbol{t}}$, εξ' ου και η ονομασία του μοντέλου ως διανυσματικού χώρου. Ορίζουμε το μέγεθος $\boldsymbol{freq}(\boldsymbol{d},\boldsymbol{t})$, ως τη **συχνότητα όρων**, δηλαδή το πόσες φορές εμφανίζεται ο όρος $\boldsymbol{t}$ στο έγγραφο $\boldsymbol{d}$. Ο πίνακας (συντελεστών βαρύτητας) των συχνοτήτων όρων, $\boldsymbol{TF}(\boldsymbol{d},\boldsymbol{t})$ καταμετράει την σχέση ενός όρου $\boldsymbol{t}$ με ένα συγκεκριμένο έγγραφο $\boldsymbol{d}$ και πρακτικά ορίζεται μηδέν (0) όταν ο όρος δεν εμφανίζεται καθόλου στο έγγραφο και διάφορο του μηδενός στις υπόλοιπες περιπτώσεις. Για αυτές τις περιπτώσεις, υπάρχουν πολλοί τρόποι να αποδοθούν οι συντελεστές βαρύτητας των συχνοτήτων όρων στο διάνυσμα $\underline{\boldsymbol{v}}$. Για παράδειγμα, μπορούμε να ορίσουμε $\boldsymbol{TF}(\boldsymbol{d},\boldsymbol{t}) = \boldsymbol{1}$, σε περίπτωση που ο όρος $\boldsymbol{t}$ εμφανίζεται στο έγγραφο $\boldsymbol{d}$, ή να χρησιμοποιήσουμε το $\boldsymbol{freq}(\boldsymbol{d},\boldsymbol{t})$, ή τη **σχετική συχνότητα όρων**, δηλαδή τη συχνότητα όρων προς τον συνολικό αριθμό συχνοτήτων όλων των όρων στο

κείμενο. Ένα άλλο παράδειγμα είναι στο σύστημα Cornell **SMART**[3], που χρησιμοποιείται ο ακόλουθος τύπος για τον υπολογισμό της κανονικοποιημένης συχνότητας όρων:

$$TF(d,t) = \begin{cases} 0 & \varepsilon\acute{\alpha}\nu\ freq(d,t) = 0 \\ 1 + \log(1 + \log(freq(d,t))) & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases}$$

Πέρα από τον όρο $\boldsymbol{freq(d,t)}$, υπάρχει άλλο ένα σημαντικό όρος ο οποίος ονομάζεται **ανάστροφη συχνότητα εγγράφου** (inverse document frequency, **IDF**) και αντιπροσωπεύει τον συντελεστή κλίμακας ή αλλιώς την σπουδαιότητα ενός όρου $\boldsymbol{t}$. Εάν ένας όρος $\boldsymbol{t}$ εμφανίζεται σε πολλά έγγραφα, η σπουδαιότητα του υποβαθμίζεται λόγω της μειωμένης διακριτικής του δύναμης. Σύμφωνα με το σύστημα Cornell **SMART**, έχουμε τον παρακάτω τύπο για την ανάστροφη συχνότητα εγγράφου:

$$IDF(t) = log\frac{1 + |d|}{|d_t|}$$

Όπου $|\boldsymbol{d}|$ είναι η συλλογή εγγράφων, και $|\boldsymbol{d_t}|$ είναι το σύνολο των εγγράφων που περιέχουν τον όρο $\boldsymbol{t}$. Αν ισχύει $|\boldsymbol{d}| \gg |\boldsymbol{d_t}|$ τότε ο όρος $\boldsymbol{t}$ έχει υψηλό συντελεστή κλίμακας και αντίστροφα.

Σε ένα ολοκληρωμένο μοντέλο διανυσματικού χώρου, οι όροι $\boldsymbol{freq(d,t)}$ και $\boldsymbol{IDF(t)}$ συνδυάζονται και σχηματίζουν τον όρο $\boldsymbol{TF - IDF(d,t)}$, ο οποίος ορίζεται ως το γινόμενο τους:

$$TF - IDF(d,t) = TF(d,t) \times IDF(t).$$

1.3.4 Τεχνικές Δεικτοδότησης Κειμένων

Δυο από τις πιο δημοφιλείς τεχνικές δεικτοδότησης κειμένων είναι αυτήν του *ανεστραμμένου καταλόγου* και των *αρχείων υπογραφής*.

Ο **ανεστραμμένος κατάλογος** είναι μια δομή καταλόγου που περιέχει δυο πίνακες κατατεμαχισμού (**hash indexed tables**) ή δυο πίνακες B+ δένδρων: τον ***πίνακα εγγράφων*** και τον ***πίνακα όρων*** όπου:

Ο ***πίνακας εγγράφων***, αποτελείται από ένα σύνολο καταγεγραμμένων εγγράφων, το καθένα απ' τα οποία περιέχει δύο πεδία: το ***doc_id*** (αναγνωριστικό εγγράφου) και το ***posting_list*** (μια λίστα από όρους ή δείκτες προς όρους, οι οποίοι εμφανίζονται στο έγγραφο, ταξινομημένοι σύμφωνα με κάποια κριτήρια σχετικότητας).

Ο ***πίνακας όρων***, αποτελείται από ένα σύνολο καταγεγραμμένων όρων, καθένας απ' τους οποίους περιέχει δύο πεδία: το ***term_id*** (αναγνωριστικό όρου) και το ***posting_list*** (μια λίστα από αναγνωριστικά εγγράφων στα οποία εμφανίζεται ο συγκεκριμένο όρος).

Με την παραπάνω μέθοδο, είναι εύκολο να απαντηθούν ερωτήματα τύπου εύρεσης εγγράφων που σχετίζονται με ένα δεδομένο σύνολο όρων ή εύρεσης όρων που σχετίζονται με ένα δεδομένο σύνολο εγγράφων. Για παράδειγμα, για να βρούμε όλα τα έγγραφα που σχετίζονται με ένα σύνολο όρων, μπορούμε πρώτα να βρούμε στον πίνακα όρων μια λίστα με αναγνωριστικά εγγράφων για κάθε όρο και μετά να τα διασταυρώσουμε ώστε να

αποκτήσουμε το σύνολο των σχετικών εγγράφων. Η μέθοδος του ανεστραμμένου καταλόγου χρησιμοποιείται ευρέως σε βιομηχανίες λόγω της ευκολίας εφαρμογής της. Οι posting lists μπορεί να είναι αρκετά μεγάλες, απαιτώντας μεγάλες αποθηκευτικές δυνατότητες, αλλά δεν είναι ικανοποιητικές όσον αφορά τον χειρισμό της ***συνωνυμίας*** (όπου δυο τελείως διαφορετικές λέξεις μπορούν να έχουν το ίδιο νόημα) και της ***πολυσημίας*** (όπου μια λέξη μπορεί να συσχετίζεται με περισσότερες από μια σημασίες).

Το **αρχείο υπογραφής** είναι ένα αρχείο που αποθηκεύει μια υπογεγραμμένη καταχώριση για κάθε έγγραφο στη βάση δεδομένων. Κάθε *υπογραφή* έχει ένα προκαθορισμένο μέγεθος από ***b*** bits που αντιπροσωπεύουν τους όρους. Ένα απλό σύστημα κωδικοποίησης έχει ως εξής: Κάθε bit της υπογραφής ενός εγγράφου αρχικοποιείται με την τιμή 0. Ένα bit παίρνει την τιμή 1 εφόσον ο όρος που αντιπροσωπεύει εμφανιστεί στο έγγραφο. Μια υπογραφή $\boldsymbol{S_1}$ταιριάζει με μια υπογραφή $\boldsymbol{S_2}$, εάν υπάρχει 1-1 αντιστοίχιση των bits των δυο υπογραφών. Επειδή συνήθως υπάρχουν περισσότεροι όροι απ' τα διαθέσιμα bits, πολλαπλοί όροι μπορεί να αντιστοιχισθούν στο ίδιο bit. Τέτοιες πολλαπλές-σε-μια αντιστοιχίσεις, μπορεί να καταστήσουν την αναζήτηση σπάταλη, διότι ένα έγγραφο που ταιριάζει στην υπογραφή ενός ερωτήματος δε σημαίνει απαραίτητα ότι περιέχει το σύνολο λέξεων-κλειδιών του ερωτήματος. Το έγγραφο πρέπει να ανακτηθεί, να αναλυθεί γραμματικά, να βρεθούν οι λεκτικοί κορμοί του και να ελεγχθεί. Για να βελτιωθεί η παραπάνω διαδικασία μπορεί να εκτελεστεί αρχικά ανάλυση συχνοτήτων, εύρεση των λεκτικών κορμών και φιλτράρισμα των stop words και στη συνέχεια να χρησιμοποιηθεί μια τεχνική κατακερματισμού και μια υπερτιθέμενη τεχνική κωδικοποίησης, για να κωδικοποιήσει το σύνολο των όρων σε αναπαράσταση bit. Παρά ταύτα, το πρόβλημα των πολλαπλών-σε-μια αντιστοιχίσεων εξακολουθεί να υπάρχει, αποτελώντας ένα από τα σημαντικότερα μειονεκτήματα αυτής της μεθόδου.

1.3.5 Τεχνικές Επεξεργασίας Ερωτημάτων

Όταν δημιουργηθεί ο ανεστραμμένος κατάλογος για μια συλλογή εγγράφων, ένα σύστημα ανάκτησης κειμένου μπορεί να «απαντήσει» γρήγορα σε ένα ερώτημα από λέξεις-κλειδιά, αναζητώντας ποια έγγραφα περιέχουν αυτές τις λέξεις. Συγκεκριμένα, συγκροτείται ένας συσσωρευτής βαθμολογίας για κάθε έγγραφο, ο οποίος ανανεώνεται κάθε φορά που διατρέχεται ο κάθε όρος του ερωτήματος. Για κάθε όρο, θα προσκομιστούν όλα τα έγγραφα που φέρουν τον όρο αυτό και θα αυξηθεί η βαθμολογία τους.

Όταν είναι διαθέσιμα παραδείγματα σχετικών εγγράφων, το σύστημα μπορεί να «μάθει» από τέτοια παραδείγματα, ώστε να βελτιώσει τα αποτελέσματα αναζήτησης. Αυτό αποκαλείται ***σχετική ανάδραση*** και έχει αποδειχθεί ότι είναι αποτελεσματικό για την βελτίωση της απόδοσης της ανάκτησης κειμένων. Σε περίπτωση που δεν είναι διαθέσιμα παραδείγματα σχετικών εγγράφων, το σύστημα μπορεί να «θεωρήσει» ως σχετικά τα πρώτα λίγα ανακτημένα έγγραφα σε κάποια αρχικά αποτελέσματα αναζητήσεων και να εξάγει περισσότερες λέξεις-κλειδιά για να διευρύνει ένα ερώτημα. Αυτή η διαδικασία καλείται ***ψευδο-ανάδραση*** ή ***τυφλή ανάδραση*** και ουσιαστικά είναι μια διαδικασία εξόρυξης χρήσιμων λέξεων-κλειδιών από τα πρώτα ανακτημένα έγγραφα στα αποτελέσματα αναζήτησης. Η ψευδο-ανάδραση οδηγεί και αυτή σε βελτιωμένα αποτελέσματα στην διαδικασία της ανάκτησης κειμένων.

Ένας βασικός περιορισμός που συναντάται στις υπάρχουσες μεθόδους ανάκτησης, είναι ότι βασίζονται στην μια-προς-μια (1-1) αντιστοίχιση των λέξεων-κλειδιών. Παρά ταύτα, λόγω της πολυπλοκότητας των φυσικών γλωσσών, αναζητήσεις βασισμένες σε λέξεις-κλειδιά, μπορεί να συναντήσουν δυο βασικά προβλήματα. Το πρώτο είναι το πρόβλημα της συνωνυμίας: δυο λέξεις με ίδιο ή παρόμοιο νόημα μπορεί να έχουν πολύ διαφορετικές επιφανειακές μορφές (για παράδειγμα στο ερώτημα ενός χρήστη μπορεί να υπάρχει η λέξη «αυτοκίνητο» αλλά ένα σχετικό κείμενο να περιέχει τη λέξη «αμάξι»). Το δεύτερο πρόβλημα που συναντάται, είναι αυτό της πολυσημίας: η ίδια λέξη μπορεί να έχει δυο διαφορετικές ερμηνείες ανάλογα με τα περιεχόμενα των εγγράφων (για παράδειγμα η λέξη «γράμμα» μπορεί να σημαίνει είτε το στοιχείο του αλφαβήτου είτε επιστολή, ανάλογα με τα περιεχόμενα του εγγράφου στο οποίο βρίσκεται).

Τα παραπάνω προβλήματα, μαζί με την μείωση του μεγέθους του ευρετηρίου, λύνονται με ανεπτυγμένες τεχνικές, οι οποίες μελετώνται στην επόμενη ενότητα.

1.3.6 Μείωση Διαστάσεων Κειμένου

Με τα κριτήρια σχετικότητας που εξετάσαμε στην παράγραφο 1.3.3 μπορούν να κατασκευαστούν ευρετήρια εγγράφων, βασισμένα στην ομοιότητα. Τα ερωτήματα που βασίζονται σε κείμενο, μπορούν να αναπαρασταθούν σαν διανύσματα, που μπορούν να χρησιμοποιηθούν για την εύρεση των κοντινών γειτόνων τους σε μια συλλογή εγγράφων. Εντούτοις, σε ειδικές βάσεις δεδομένων εγγράφων, ο αριθμός των όρων $\boldsymbol{T}$ και ο αριθμός των εγγράφων $\boldsymbol{D}$ είναι συνήθως αρκετά μεγάλος. Κατά συνέπεια ο πίνακας σχετικότητας θα έχει μέγεθος $\boldsymbol{T} \times \boldsymbol{D}$, οδηγώντας σε υψηλή διαστατικότητα και κατά συνέπεια σε ανεπαρκή αναζήτηση, αραιά διανύσματα και αυξημένη δυσκολία στην ανακάλυψη και εκμετάλλευση σχέσεων μεταξύ των όρων (πχ. συνωνυμία). Για να ξεπεραστούν αυτά τα προβλήματα, χρησιμοποιούνται τεχνικές μείωσης της διαστατικότητας όπως η ***Λανθάνουσα Σημασιολογική Δεικτοδότηση*** (Latent Semantic Indexing, **LSI**), η ***Πιθανοθεωρητική Λανθάνουσα Σημασιολογική Ανάλυση*** (Probabilistic Latent Semantic Analysis, **PLSA**) και η ***Δεικτοδότηση Διατήρησης Τοπικότητας*** (Locality Preserving Indexing, **LPI**).

Στη συνέχεια, για να εξηγηθεί καλύτερα η βασική ιδέα πίσω από τις δυο κυριότερες μεθόδους, την **LSI** και την **LPI**, θα θεωρήσουμε τους παρακάτω ορισμούς:

$\boldsymbol{x_1}, \dots, \boldsymbol{x_{tn}} \in \mathbb{R}^{\boldsymbol{m}}$**,** διανύσματα που αντιπροσωπεύουν $\boldsymbol{n}$ έγγραφα με $\boldsymbol{m}$ λέξεις.

$\boldsymbol{X} = [\boldsymbol{x_1}, \dots, \boldsymbol{x_{tn}}]$**,** πίνακας που περιέχει τα παραπάνω διανύσματα.

1.3.6.1 Λανθάνουσα Σημασιολογική Δεικτοδότηση (LSI)

Η ***Λανθάνουσα Σημασιολογική Δεικτοδότηση*** (από εδώ και πέρα, **LSI**) είναι ένας από τους δημοφιλέστερους αλγόριθμους που χρησιμοποιείται για την μείωση της διαστατικότητας των εγγράφων. Είναι θεμελιωδώς βασισμένος στην ***Διάσπαση Ιδιαζουσών Τιμών*** (Singular Value Decomposition: **SVD**). Σύμφωνα με αυτήν, κάθε μητρώο μπορεί να μετατραπεί σε διαγώνιο με πολλαπλασιασμό τόσο από αριστερά, αλλά και από δεξιά με κατάλληλα ορθομοναδιαία μητρώα. Πιο συγκεκριμένα έχουμε τον παρακάτω ορισμό:

Ορισμός. *Έστω* $\mathbb{A} \in \mathbb{C}^{m\times n}$. *Υπάρχουν δυο ορθομοναδιαία μητρώα* $\mathbb{U} \in \mathbb{C}^{m\times m}$ *και* $\mathbb{V} \in \mathbb{C}^{n\times n}$, *τέτοια ώστε:*

$$\mathbb{U}^*\mathbb{A}\mathbb{V} = \boldsymbol{\Sigma} = diag(\sigma_1, \dots, \sigma_p),\ p = min(m, n) \qquad (1)$$

και $\sigma_1 \geq \cdots \geq \sigma_p \geq 0$. *Η (1) καλείται Διάσπαση Ιδιαζουσών Τιμών ή SVD του* $\mathbb{A}$ *και οι αριθμοί* σ_i ή $\sigma_i(\mathbb{A})$ *καλούνται ιδιάζουσες τιμές του* $\mathbb{A}$.

Από τον παραπάνω ορισμό, προκύπτει επιπλέον ότι η ιδιάζουσα παραγοντοποίηση του μητρώου $\mathbb{A}$, δίνεται από την παρακάτω σχέση:

$$\mathbb{A} = \mathbb{U}\boldsymbol{\Sigma}\mathbb{V}^T \in \mathbb{C}^{m\times n}$$

Υποθέτοντας ότι ο βαθμός του πίνακα $\boldsymbol{X}$, που ορίστηκε παραπάνω, είναι $r = rank[\boldsymbol{X}]$, η LSI αποσυνθέτει τον πίνακα $\boldsymbol{X}$ με SVD ως εξής:

$$\boldsymbol{X} = \mathbb{U}\boldsymbol{\Sigma}\mathbb{V}^T$$

όπου $\boldsymbol{\Sigma} = diag(\sigma_1, \dots, \sigma_r)$ και $\sigma_1 \geq \cdots \geq \sigma_r$. Η μέθοδος **LSI** χρησιμοποιεί τα πρώτα *k* διανύσματα του μητρώου $\mathbb{U}$ ως τον πίνακα μετασχηματισμού για να ενσωματώσει τα έγγραφα σε ένα k-διάστατο υποχώρο. Εύκολα μπορεί να ελεγχθεί ότι τα διανύσματα στις στήλες του $\mathbb{U}$ είναι οι ιδιοτιμές του $\boldsymbol{X}\boldsymbol{X}^T$. Η βασική ιδέα στην μέθοδο **LSI** είναι να εξάγει τα πιο αντιπροσωπευτικά χαρακτηριστικά και ταυτόχρονα να ελαχιστοποιήσει το σφάλμα αναδόμησης. Έστω $\boldsymbol{a}$ το διάνυσμα μετασχηματισμού. Η βασική συνάρτηση της μεθόδου **LSI** μπορεί τότε να διατυπωθεί ως εξής:

$$\boldsymbol{a_{opt}} = arg\min_{\boldsymbol{a}} \|\boldsymbol{X} - \boldsymbol{a}\boldsymbol{a}^T\boldsymbol{X}\|^2 = arg\max_{\boldsymbol{a}} \boldsymbol{a}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{a}$$

με τον περιορισμό

$$\boldsymbol{a}^T\boldsymbol{a} = 1.$$

Εφόσον ο πίνακας $\boldsymbol{X}\boldsymbol{X}^T$είναι συμμετρικός, οι βασικές συναρτήσεις της μεθόδου **LSI** είναι ορθογώνιες.

1.3.6.2 Δεικτοδότηση Διατήρησης Τοπικότητας (LPI)

Η μέθοδος ***Δεικτοδότησης Διατήρησης Τοπικότητας*** (από δω και πέρα, **LPI**) στοχεύει στην εξαγωγή των πιο διακριτικών στοιχείων σε αντίθεση με την **LSI** που στοχεύει στην εξαγωγή των πιο αντιπροσωπευτικών στοιχείων. Η βασική ιδέα της **LPI**, είναι να διατηρήσει την τοπικότητα της πληροφορίας (για παράδειγμα, εάν δυο έγγραφα «γειτνιάζουν» στον αρχικό χώρο που ορίζουν τα έγγραφα, η **LPI** θα προσπαθήσει να τα «κρατήσει» κοντά, όταν μειωθεί διαστατικά ο χώρος). Από τη στιγμή που τα γειτονικά κείμενα (δείκτες δεδομένων στον υψηλών-διαστάσεων χώρο) έχουν πιθανώς κοινό θέμα, η **LPI** μπορεί να χαρτογραφήσει όσο πιο κοντά το ένα στο άλλο, τα κείμενα που είναι σημασιολογικά σχετικά μεταξύ τους.

Όπως και στην προηγούμενη μέθοδο, θεωρούμε $\boldsymbol{x_1}, \dots, \boldsymbol{x_{tn}} \in \mathbb{R}^{\boldsymbol{m}}$, διανύσματα που αντιπροσωπεύουν $\boldsymbol{n}$ έγγραφα με $\boldsymbol{m}$ λέξεις. Δεδομένου ενός πίνακα ομοιότητας $\boldsymbol{S} \in \mathbb{R}^{n\times n}$, η **LPI** μπορεί να προσδιοριστεί από την επίλυση του παρακάτω προβλήματος ελαχιστοποίησης:

$$\boldsymbol{a_{opt}} = arg\min_{\boldsymbol{a}} \sum_{i=1}^{n} (\boldsymbol{a}^T \boldsymbol{x}_i - \boldsymbol{a}^T \boldsymbol{x}_j)^2 \boldsymbol{S}_{ij} = arg\min_{\boldsymbol{a}} \boldsymbol{a}^T \boldsymbol{X} \boldsymbol{L} \boldsymbol{X}^T \boldsymbol{a}$$

με τον περιορισμό:

$$\boldsymbol{a}^T \boldsymbol{X} \boldsymbol{D} \boldsymbol{X}^T = 1.$$

όπου $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S}$ είναι η Λαπλασιανή μήτρα του πίνακα $\boldsymbol{S}$ και τα στοιχεία $\boldsymbol{D_{ii}} = \sum_{\boldsymbol{j}} \boldsymbol{S}_{ij}$, μετράνε την τοπική πυκνότητα γύρω από το $\boldsymbol{x}_i$. Η «κατασκευή» του πίνακα $\boldsymbol{S}$ γίνεται ως εξής:

$$\boldsymbol{S}_{ij} = \begin{cases} \boldsymbol{x}_i^T \boldsymbol{x}_j, & \text{αν το } \boldsymbol{x}_i \text{ είναι ανάμεσα στους } p \text{ γείτονες του } \boldsymbol{x}_j \text{ (ή το αντίστροφο)} \\ 0, & \text{αλλού} \end{cases}$$

Έτσι, η αντικειμενική συνάρτηση της **LPI**, επισύρει βαριά ποινή εάν τα γειτονικά σημεία $\boldsymbol{x}_i$και $\boldsymbol{x}_j$ είναι καθορισμένα μακριά το ένα από το άλλο. Κατ' αυτόν τον τρόπο, ελαχιστοποιώντας τη συνάρτηση αυτή, μπορεί να εξασφαλιστεί ότι, αν τα $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ είναι «κοντά» μεταξύ τους, τότε οι εικόνες τους, $y_i (= \boldsymbol{a}^T \boldsymbol{x}_i)$ και $y_j (= \boldsymbol{a}^T \boldsymbol{x}_j)$ είναι και αυτές «κοντά» μεταξύ τους. Ολοκληρώνοντας, οι βασικές συναρτήσεις της **LPI** είναι τα ιδιοδιανύσματα που σχετίζονται με τις μικρότερες ιδιοτιμές του παρακάτω γενικού προβλήματος ιδιοτιμών:

$$\boldsymbol{X} \boldsymbol{L} \boldsymbol{X}^T \boldsymbol{a} = \lambda \boldsymbol{X} \boldsymbol{D} \boldsymbol{X}^T \boldsymbol{\alpha}.$$

Συγκρίνοντας τις μεθόδους **LSI** και **LPI** καταλήγουμε στα παρακάτω:

Η μέθοδος **LSI** στοχεύει στον εντοπισμό της καλύτερης προσέγγισης ενός υπο-χώρου που προέρχεται από τον αρχικό χώρο που ορίζουν τα έγγραφα, υπό την έννοια της ελαχιστοποίησης του ολικού λάθους ανακατασκευής. Με άλλα λόγια, η **LSI** αναζητά τα πιο χαρακτηριστικά στοιχεία. Η μέθοδος **LPI** στοχεύει στον εντοπισμό της τοπικής γεωμετρικής δομής του χώρου που ορίζουν τα έγγραφα. Από τη στιγμή που γειτονικά έγγραφα σχετίζονται όσον αφορά το θέμα, η μέθοδος αυτή έχει περισσότερη διακριτική ικανότητα σε σχέση με την **LSI**. Από θεωρητική ανάλυση της **LPI** προκύπτει ότι είναι μια μη ελεγχόμενη προσέγγιση της ελεγχόμενης ***Γραμμικής Διακριτικής Ανάλυσης*** (Linear Discriminant Analysis: **LDA**) και κατά συνέπεια για κατηγοριοποίηση και περισυλλογή κειμένων αναμένεται να έχει καλύτερα αποτελέσματα σε σχέση με την **LSI**.

1.3.6.3 Πιθανοθεωρητική Λανθάνουσα Σημασιολογική Δεικτοδότηση (PLSI)

Η ***Πιθανοθεωρητική Λανθάνουσα Σημασιολογική Δεικτοδότηση*** (από δω και πέρα, **PLSI**) είναι μια παρόμοια μέθοδος με την **LSI**, αλλά επιτυγχάνει μείωση της διαστατικότητας με ένα πιθανοθεωρητικό μοντέλο μίξης.

Πιο συγκεκριμένα, θεωρούμε ότι υπάρχουν $\boldsymbol{k}$ λανθάνοντα κοινά θέματα σε μια συλλογή εγγράφων και καθένα από αυτά χαρακτηρίζεται από μια πολυωνυμική κατανομή λέξεων. Ένα έγγραφο λαμβάνεται ως ένα δείγμα από μοντέλο μίξης με τα παραπάνω μοντέλα θεμάτων ως συνιστώσες. Το μοντέλο μίξης «ταιριάζεται» σε όλα στα έγγραφα και τα

αποκτώμενα $\boldsymbol{k}$ πολυωνυμικά μοντέλα συνιστωσών μπορεί να θεωρηθεί ότι ορίζουν $\boldsymbol{k}$ νέες σημασιολογικές διαστάσεις. Τα βάρη μίξης ενός εγγράφου μπορούν να χρησιμοποιηθούν για μια νέα αναπαράσταση του ίδιου του εγγράφου στις μικρότερες λανθάνουσες σημασιολογικές διαστάσεις.

Θεωρούμε $\boldsymbol{C} = [\boldsymbol{d_1}, \dots, \boldsymbol{d_n}]$ μια συλλογή $\boldsymbol{n}$ εγγράφων και $\boldsymbol{\theta_1}, \dots, \boldsymbol{\theta_k}$, $\boldsymbol{k}$ πολυωνυμικές κατανομές θεμάτων. Μια λέξη $\boldsymbol{w}$ σε ένα έγγραφο $\boldsymbol{d_i}$ λαμβάνεται ως δείγμα του ακόλουθου μοντέλου μίξης:

$$p_{d_i}(w) = \sum_{j=1}^{k} [\pi_{d_i,j} p(w|\theta_j]$$

όπου οι όροι $\boldsymbol{\pi_{d_i,j}}$ αποτελούν το βάρος μίξης ενός καθορισμένου εγγράφου $\boldsymbol{d_i}$για το $\boldsymbol{j}$ -οστό θέμα και για τα οποία ισχύει η παρακάτω σχέση:

$$\sum_{j=1}^{k} \pi_{d_i,j} = 1$$

Η λογαριθμική πιθανότητα της συλλογής $\boldsymbol{C}$ είναι:

$$\log p(C|\boldsymbol{\Lambda}) = \sum_{i=1}^{n} \sum_{w \in \mathbb{V}} [c(w, d_i) \log(\sum_{j=1}^{k} [\pi_{d_i,j} p(w|\theta_j])]$$

όπου:

$\mathbb{V}$ είναι το σύνολο όλων των λέξεων (πχ. λεξιλόγιο)

$c(w, d_i)$ είναι το πόσες φορές εμφανίζεται η λέξη $\boldsymbol{w}$ στο κείμενο $\boldsymbol{d_i}$

$\boldsymbol{\Lambda} = (\{\theta_j, \{\pi_{d_i,j}\}_{i=1}^{n}\}_{j=1}^{k})$ είναι το σύνολο όλων των παραμέτρων των μοντέλων των θεμάτων.

Το παραπάνω μοντέλο μπορεί να εκτιμηθεί χρησιμοποιώντας τον αλγόριθμο ***Προσδοκίας-Μεγιστοποίησης*** (Expectation – Maximization: **EM**) που υπολογίζει την παρακάτω μέγιστη εκτίμησης πιθανότητας:

$$\widehat{\boldsymbol{\Lambda}} = \arg\max_{\boldsymbol{\Lambda}} \log p(C|\boldsymbol{\Lambda})$$

Όταν γίνει η εκτίμηση, οι $\boldsymbol{\theta_1}, \dots, \boldsymbol{\theta_k}$, ορίζουν k νέες σημασιολογικές διαστάσεις και τα βάρη $\pi_{d_i,j}$ δίνουν μια νέα αναπαράσταση των εγγράφων d_i, στον νέο, μικρότερων διαστάσεων χώρο.

1.3.7 Προσεγγίσεις Εξόρυξης Δεδομένων από Κείμενα

Υπάρχουν πολλές προσεγγίσεις στην εξόρυξη δεδομένων από κείμενα οι οποίες μπορούν να ταξινομηθούν από διάφορες προοπτικές, με βάση τις εισόδους που λαμβάνονται σε ένα σύστημα εξόρυξης δεδομένων και τις διεργασίες που πρόκειται να εκτελεστούν. Σε γενικές γραμμές, οι σημαντικές προσεγγίσεις με βάση τα είδη των δεδομένων που λαμβάνονται ως είσοδος, είναι οι εξής:

- η προσέγγιση που βασίζεται στη *λέξη-κλειδί*, όπου η είσοδος είναι ένα σύνολο από λέξεις-κλειδιά ή όρους στα έγγραφα,
- η προσέγγιση της *ετικετοποίησης*, όπου η είσοδος είναι ένα σύνολο από ετικέτες, και
- η προσέγγιση της *εξαγωγής πληροφοριών*, όπου οι εισροές είναι σημασιολογικές πληροφορίες, όπως γεγονότα, δεδομένα ή οντότητες που έχουν αποκαλυφθεί από εξαγωγή πληροφοριών.

Μια απλή προσέγγιση που θα βασίζεται σε μια *λέξη-κλειδί*, μπορεί μόνο να ανακαλύψει σχέσεις σε σχετικά επιφανειακό επίπεδο, όπως επανακάλυψη σύνθετων ουσιαστικών (πχ. "βάση δεδομένων" και "συστήματα") ή επανεμφανιζόμενα μοτίβα με μικρότερη σημασία (πχ. "τρομοκράτης" και "έκρηξη"). Δεν μπορεί να φέρει πολύ βαθιά κατανόηση του κειμένου. Η προσέγγιση της *ετικετοποίησης* μπορεί να βασιστεί σε ετικέτες που λαμβάνονται με χειροκίνητη τοποθέτηση ετικετών (μια διαδικασία που είναι δαπανηρή και ανέφικτη για μεγάλες συλλογές εγγράφων) ή από κάποιον αλγόριθμο αυτόματης κατηγοριοποίησης (που μπορεί να επεξεργαστεί ένα σχετικά μικρό σύνολο ετικετών και απαιτεί προκαταβολικό προσδιορισμό των διαφόρων κατηγοριών). Η προσέγγιση της *εξαγωγής πληροφοριών* είναι πιο προηγμένη και μπορεί να οδηγήσει στην ανακάλυψη κάποιας βαθύτερης γνώσης, αλλά απαιτεί σημασιολογική ανάλυση του κειμένου με κατανόηση φυσικής γλώσσας και μεθόδους μάθησης μηχανής, θέτοντας απαιτητικούς στόχους για την ανακάλυψη γνώσης.

Διάφορες εργασίες πάνω στην εξόρυξη κειμένου μπορούν να εκτελεστούν στις εξαγόμενες λέξεις-κλειδιά, ετικέτες ή σημασιολογικές πληροφορίες. Αυτές συμπεριλαμβάνουν την συσταδοποίηση (clustering) εγγράφων, την ταξινόμηση, την άντληση πληροφοριών καθώς και την ανάλυση σχέσεων και τάσεων.

1.3.7.1 Σχεσιακή Ανάλυση Βασισμένη σε Λέξεις-Κλειδιά

Η ανάλυση αυτή συλλέγει σύνολα λέξεων-κλειδιών ή όρων που επαναλαμβάνονται συχνά μαζί και βρίσκει τις σχέσεις ή τους συσχετισμούς μεταξύ τους. Όπως οι περισσότερες αναλύσεις σε βάσεις δεδομένων κειμένων, έτσι και η σχεσιακή ανάλυση αρχικά προ-επεξεργάζεται τα δεδομένα μέσω ανάλυσης, εύρεσης του λεκτικού κορμού τους, αφαίρεσης των stop words και άλλων τεχνικών. Στη συνέχεια καλεί αλγόριθμους εξόρυξης σχέσεων/συσχετίσεων. Σε μια βάση δεδομένων εγγράφων, κάθε έγγραφο μπορεί να εκληφθεί σαν μια συναλλαγή, ενώ ένα σύνολο λέξεων-κλειδιών στο κείμενο μπορεί να θεωρηθεί σαν ένα σύνολο όρων για την διεκπεραίωση της συναλλαγής. Με την παραπάνω λογική, η βάση δεδομένων είναι στην παρακάτω μορφή:

$$ΒΔ = \{αναγνωριστικό\_εγγράφου, σύνολο\_λέξεων\_κλειδιών\}$$

Με αυτόν τον τρόπο, το πρόβλημα της εξόρυξης σχέσεων βασισμένων σε λέξεις-κλειδιά σε βάσεις δεδομένων εγγράφων ανάγεται στην εξόρυξη σχέσεων σε αντικείμενα σε βάσεις δεδομένων συναλλαγών.

Πρέπει να σημειωθεί ότι ένα σύνολο συχνά εμφανιζόμενων διαδοχικών ή κοντά τοποθετημένων λέξεων-κλειδιών μπορεί να σχηματίζει έναν *όρο* ή μια *φράση*. Η σχεσιακή ανάλυση μπορεί να βοηθήσει στον εντοπισμό *σύνθετων σχέσεων* (σχέσεις που δηλώνουν κάποια μορφή κτήσης, όπως για παράδειγμα [*Μετσόβιο*, *Πολυτεχνείο*] ή [*Ελλάδα*, *Πρόεδρος*, *Κάρολος Παπούλιας*]) ή *απλών σχέσεων* (για παράδειγμα [*δολάριο*, *ευρώ*, *συνάλλαγμα*,

μετοχές, *υπόλοιπο*, *επιτροπή*, *ασφάλεια*]). Η εξόρυξη που βασίζεται σε αυτών των ειδών τις σχέσεις αναφέρεται ως «**εξόρυξη σχέσεων σε επίπεδο όρων**» (σε αντίθεση με την εξόρυξη που βασίζεται σε μεμονωμένες λέξεις). Η αναγνώριση όρων και η εξόρυξη συσχετίσεων σε επίπεδο όρων έχει δυο πλεονεκτήματα στην ανάλυση κειμένων:

- Οι όροι και οι φράσεις ετικετοποιούνται αυτόματα και δεν υπάρχει έτσι ανάγκη για ανθρώπινη παρέμβαση στην ετικετοποίηση εγγράφων, και
- Ο αριθμός των άστοχων αποτελεσμάτων μειώνεται ραγδαία, όπως και ο χρόνος εκτέλεσης των αλγορίθμων εξόρυξης.

Με την αναγνώριση όρων και φράσεων, η εξόρυξη σε επίπεδο όρων μπορεί να κληθεί για να βρει σχέσεις ανάμεσα σε ένα σύνολο εντοπισμένων όρων και λέξεων-κλειδιών. Ορισμένοι χρήστες μπορεί να θέλουν να βρουν σχέσεις ανάμεσα σε ζευγάρια λέξεων-κλειδιών ή όρων από ένα δεδομένο σύνολο λέξεων-κλειδιών ή φράσεων, ενώ άλλοι χρήστες μπορεί να θέλουν να βρουν το μέγιστο σύνολο όρων που εμφανίζονται μαζί. Επομένως, ανάλογα με τις απαιτήσεις εξόρυξης των χρηστών, μπορούν να κληθούν βασικοί σχεσιακοί αλγόριθμοι εξόρυξης ή αλγόριθμοι μεγιστοποίησης προτύπων.

1.3.7.2 Ανάλυση Ταξινόμησης Εγγράφων

Η αυτοματοποιημένη ταξινόμηση εγγράφων είναι μια σημαντική διεργασία της εξόρυξης κειμένων, επειδή με την ύπαρξη ενός τεράστιου αριθμού online εγγράφων, είναι κουραστικό αλλά συνάμα απαραίτητο να οργανώσει αυτόματα τα διάφορα αυτά έγγραφα σε κατηγορίες για τη διευκόλυνση της ανάκτησης και μετέπειτα ανάλυσης των. Η ταξινόμηση εγγράφων έχει χρησιμοποιηθεί στην αυτοματοποιημένη ετικετοποίηση θεμάτων (δηλαδή, αντιστοίχιση ετικετών σε έγγραφα), στην κατασκευή ευρετηρίων θεμάτων, στην ταυτοποίηση εγγράφων βάσει του τρόπου γραφής των (η οποία μπορεί να βοηθήσει στον εντοπισμό των συγγραφέων ανώνυμων εγγράφων) και στην ταξινόμηση των σκοπών των υπερ-συνδέσεων που συνδέονται με ένα σύνολο εγγράφων.

Η γενική διαδικασία περιλαμβάνει τα παρακάτω: αρχικά, ένα σύνολο αταξινόμητων εγγράφων λαμβάνεται ως το *δοκιμαστικό σύνολο*. Το σύνολο αυτό αναλύεται στη συνέχεια προκειμένου να παραχθεί ένα σύστημα ταξινόμησης. Ένα τέτοιο σύστημα ταξινόμησης συχνά πρέπει να τελειοποιηθεί με μια διαδικασία δοκιμής. Το σύστημα ταξινόμησης που παράγεται με αυτόν τον τρόπο μπορεί στη συνέχεια να χρησιμοποιηθεί για την ταξινόμηση των άλλων online εγγράφων.

Αυτή η διαδικασία φαίνεται παρόμοια με την ταξινόμηση των σχεσιακών δεδομένων. Ωστόσο υπάρχει μια θεμελιώδης διαφορά. Τα σχεσιακά δεδομένα είναι καλά δομημένα: κάθε πλειάδα δεδομένων ορίζεται από ένα ζεύγος χαρακτηριστικών-τιμών. Για παράδειγμα, στην πλειάδα *{ηλιόλουστος, ζέστη, στεγνό, αέρας, παιχνίδι, τένις}*, η τιμή "*ηλιόλουστος*" αντιστοιχεί στο χαρακτηριστικό *καιρός*, η τιμή "*ζέστη*" αντιστοιχεί στο χαρακτηριστικό *θερμοκρασία*, και ούτω καθεξής. Η ανάλυση ταξινόμησης αποφασίζει ποια σύνολα από ζεύγη χαρακτηριστικών-τιμών έχουν την υψηλότερη διακριτική δύναμη στον καθορισμό του «*αν ένα άτομο πρόκειται να παίξει τένις*». Από την άλλη πλευρά, οι βάσεις δεδομένων εγγράφων δεν είναι δομημένες σύμφωνα με ζεύγη χαρακτηριστικών-τιμών. Δηλαδή, ένα σύνολο λέξεων-κλειδιών που σχετίζεται με ένα σύνολο εγγράφων δεν είναι οργανωμένο σε ένα σταθερό σύνολο χαρακτηριστικών ή διαστάσεων. Αν αντιμετωπιστεί ξεχωριστά κάθε

λέξη-κλειδί, όρος ή χαρακτηριστικό στο έγγραφο ως μια διάσταση, μπορεί να υπάρξουν χιλιάδες διαστάσεις σε ένα σύνολο των εγγράφων. Ως εκ τούτου, η χρησιμοποίηση συνηθισμένων σχεσιακών μεθόδων ταξινόμησης που εστιάζουν σε δεδομένα, όπως η ανάλυση με δέντρα αποφάσεων, μπορεί να μην είναι αποτελεσματική για την ταξινόμηση των βάσεων δεδομένων εγγράφων.

Μερικές τυπικές μέθοδοι ταξινόμησης που έχουν χρησιμοποιηθεί με επιτυχία στην ταξινόμηση κειμένων είναι η ***ταξινόμηση πλησιέστερου γείτονα*** (nearest-neighbor classification), η μέθοδος ***επιλογής χαρακτηριστικών*** (feature selection method), η ***Bayesian ταξινόμηση***, οι ***μηχανές διανυσμάτων υποστήριξης*** (support vector machines) και η ***σχεσιακή ταξινόμηση*** (association-based).

Σύμφωνα με το μοντέλο Διανυσματικού Χώρου, δύο έγγραφα είναι παρόμοια αν μοιράζονται παρόμοια διανύσματα εγγράφων. Αυτό το μοντέλο παρακινεί την κατασκευή ***ταξινομητή βάσει του k-πλησιέστερου-γείτονα***, λόγω της «διαίσθησης» ότι σε παρόμοια έγγραφα αναμένεται να ανατεθεί ετικέτα της ίδιας τάξης. Όλα τα δοκιμαστικά έγγραφα επομένως μπορούν να ευρετηριοποιηθούν, το καθένα με την αντίστοιχη ετικέτα κατηγορίας. Όταν υποβληθεί ένα έγγραφο δοκιμαστικά, αυτό μπορεί να αντιμετωπιστεί ως ερώτημα στο σύστημα Ανάκτησης Πληροφοριών και να ανακτηθούν $\boldsymbol{k}$ έγγραφα από το σύνολο των δοκιμαστικών εγγράφων, που είναι τα πιο σχετικά με το ερώτημα (όπου $\boldsymbol{k}$ είναι μια ρυθμιζόμενη σταθερά). Η ετικέτα της κατηγορίας που ανήκει το δοκιμαστικό έγγραφο μπορεί να καθοριστεί βάσει της ετικέτας κατηγορίας που έχει διανεμηθεί στους $\boldsymbol{k}$ πλησιέστερους γείτονες. Η διανομή της ετικέτας κατηγορίας μπορεί επιπλέον να βελτιωθεί, όπως για παράδειγμα να βασιστεί στις σταθμισμένες μετρήσεις αντί για τις πρώτες, ή αφήνοντας κατά μέρος ένα τμήμα των κατηγοριοποιημένων εγγράφων για επικύρωση. Ρυθμίζοντας το $\boldsymbol{k}$ και ενσωματώνοντας τις προτεινόμενες βελτιώσεις, αυτό το είδος ταξινομητή μπορεί να επιτύχει ακρίβεια συγκρίσιμη με αυτή του καλύτερου ταξινομητή. Ωστόσο, δεδομένου ότι η μέθοδος αυτή χρειάζεται μη τετριμμένο χώρο (ενδεχομένως πλεονάζοντα) για αποθήκευση δοκιμαστικών πληροφοριών και επιπλέον χρόνο για αναζήτηση ανεστραμμένου ευρετηρίου, υστερεί σε επιπλέον χώρο και χρόνο γενικά σε σύγκριση με άλλα είδη ταξινομητών.

Το μοντέλο Διανυσματικού Χώρου μπορεί να αντιστοιχίσει μεγάλο «βάρος» σε σπάνια αντικείμενα, ανεξάρτητα από τα χαρακτηριστικά της κατηγορίας στην οποία ανήκουν. Τέτοια σπάνια αντικείμενα μπορεί να οδηγήσουν σε μη αποτελεσματική ταξινόμηση. Αυτό γίνεται καλύτερα κατανοητό μέσα από το παρακάτω παράδειγμα υπολογισμού του μέτρου *TF-IDF*:

Παράδειγμα : Έστω ότι υπάρχουν δυο όροι t_1, t_2 σε δύο κλάσεις C_1, C_2, που η καθεμία περιέχει 100 δοκιμαστικά έγγραφα. Ο πρώτος όρος t_1, εμφανίζεται σε 5 έγγραφα στην κάθε κλάση (δηλαδή στο 5% του συνολικού όγκου των εγγράφων) και ο δεύτερος όρος t_2, σε 20 έγγραφα στην κλάση C_1 μόνο (δηλαδή στο 10% του συνολικού όγκου των εγγράφων). Ο όρος t_1 θα έχει υψηλότερο δείκτη *TF-IDF*, επειδή είναι πιο σπάνιος, αλλά είναι προφανές ότι ο όρος t_2 έχει υψηλότερη διακριτική δύναμη στην παραπάνω περίπτωση. Μια διαδικασία ***επιλογής χαρακτηριστικών*** μπορεί να χρησιμοποιηθεί ώστε να αφαιρέσει τους όρους, που είναι στατιστικά ασύνδετοι με τις ετικέτες κατηγοριών, από τα δοκιμαστικά έγγραφα. Με αυτόν τον τρόπο θα μειωθεί το σύνολο των όρων που χρησιμοποιούνται για ταξινόμηση, βελτιώνοντας έτσι τόσο την αποτελεσματικότητα όσο και την ακρίβεια.

Μετά την διαδικασία επιλογής χαρακτηριστικών, η οποία αφαιρεί τους μη χαρακτηριστικούς όρους, τα εναπομείναντα «εξυγιασμένα» δοκιμαστικά κείμενα μπορούν

να χρησιμοποιηθούν για μια αποτελεσματική ταξινόμηση. Η ***Bayesian ταξινόμηση*** είναι μια από τις πολλές δημοφιλείς τεχνικές που μπορεί να χρησιμοποιηθεί για την αποτελεσματική ταξινόμηση εγγράφων. Δεδομένου ότι η ταξινόμηση εγγράφων μπορεί να θεωρηθεί ως ο υπολογισμός της στατιστικής κατανομής των εγγράφων σε συγκεκριμένες κατηγορίες, ένας *Bayesian ταξινομητής* «εκπαιδεύει» αρχικά το μοντέλο, υπολογίζοντας μια γενικευμένη κατανομή εγγράφων $\boldsymbol{P(d|c)}$ για κάθε κατηγορία $\boldsymbol{c}$ του εγγράφου $\boldsymbol{d}$ και στη συνέχεια «δοκιμάζει» ποια κατηγορία είναι πιθανότερο να δημιουργήσει το δοκιμαστικό έγγραφο.

Επειδή οι παραπάνω δυο μέθοδοι χειρίζονται υψηλών διαστάσεων σύνολα δεδομένων, μπορούν να χρησιμοποιηθούν και οι δυο για αποτελεσματική ταξινόμηση εγγράφων. Άλλοι μέθοδοι ταξινόμησης όπως η μέθοδος ***γραμμικής παλινδρόμησης ελαχίστων τετραγώνων*** και ***μηχανές διανυσμάτων υποστήριξης*** (**support vector machines**) μπορούν να χρησιμοποιηθούν για αποτελεσματική κατηγοριοποίηση.

Η τελευταία από τις αναφερθείσες μεθόδους είναι η **σχεσιακή ταξινόμηση**, η οποία ταξινομεί τα έγγραφα βάσει ενός συνόλου αλληλοσχετιζόμενων και συχνά εμφανιζόμενων μοτίβων. Συχνά εμφανιζόμενοι όροι όμως, όπως έχει ήδη σημειωθεί, δεν έχουν ισχυρή διακριτική δύναμη, με αποτέλεσμα μόνο οι όροι που δεν εμφανίζονται τόσο συχνά και ως εκ τούτου έχουν καλή διακριτική δύναμη μπορούν να χρησιμοποιηθούν για την ταξινόμηση των κειμένων. Η διαδικασία που ακολουθείται κατά την εφαρμογή της παραπάνω μεθόδου είναι η ακόλουθη:

Αρχικά, διάφορες λέξεις-κλειδιά και όροι μπορούν να εξαχθούν από την ανάκτηση πληροφοριών και από απλές σχεσιακές μεθόδους ανάλυσης. Στη συνέχεια, μπορούν να επιτευχθούν εννοιολογικές ιεραρχίες των λέξεων-κλειδιών και των όρων, με τη χρήση διαθέσιμων τάξεων όρων, όπως το **Wordnet**[4], ή βασισμένες στη γνώση ειδικών ή από συστήματα ταξινόμησης λέξεων-κλειδιών. Τα έγγραφα που περιλαμβάνονται στο δοκιμαστικό σύνολο μπορούν και αυτά να ταξινομηθούν σε ιεραρχίες. Μια μέθοδος εξόρυξης σχέσεων όρων εφαρμόζεται στη συνέχεια, ώστε να ανακαλυφθούν σύνολα αλληλοσχετιζόμενων όρων που μπορούν να χρησιμοποιηθούν για να διαχωρίσουν μια τάξη κειμένων από μια άλλη. Η διαδικασία αυτή παράγει ένα σύνολο κανόνων συσχέτισης που συνδέονται με κάθε τάξη κειμένων. Οι κανόνες αυτοί μπορούν να κανονιστούν ανάλογα με την διακριτική τους δύναμη και τη συχνότητα εμφάνισης και να χρησιμοποιηθούν για να ταξινομήσουν νέα έγγραφα. Τέτοιου είδους σχεσιακοί ταξινομητές εγγράφων, έχουν αποδειχθεί ιδιαίτερα αποτελεσματικοί.

1.3.7.3 Ανάλυση Συσταδοποίησης Κειμένων

Η ***συσταδοποίηση εγγράφων*** (document clustering) είναι μια από τις πιο κρίσιμες τεχνικές για την οργάνωση εγγράφων υπό μη ελεγχόμενο τρόπο. Όταν τα έγγραφα αντιπροσωπεύονται από όρους-διανύσματα, συγκεκριμένες μέθοδοι όπως ιεραρχικοί αλγόριθμοι συσταδοποίησης και ο K-means αλγόριθμος, μπορούν να εφαρμοστούν. Εξαιτίας όμως της πολύ υψηλής διαστατικότητας του χώρου των εγγράφων, της λεγόμενης «*κατάρας της διαστατικότητας*», είναι λογικό να προβληθούν πρώτα τα έγγραφα σε έναν χαμηλότερων διαστάσεων υποχώρο, στον οποίο η σημασιολογική δομή του αρχικού χώρου να είναι ξεκάθαρη. Σ' αυτόν τον χαμηλής διαστατικότητας σημασιολογικό χώρο, οι παραδοσιακοί αλγόριθμοι συσταδοποίησης μπορούν να εφαρμοστούν. Για την επίτευξη των ανωτέρω

χρησιμοποιούνται μέθοδοι, όπως η ***φασματική συσταδοποίηση*** (spectral clustering), ***μίγμα μοντέλων συσταδοποίησης***, καθώς και η ***συσταδοποίηση με χρήση των μεθόδων LSI και LPI***.

Η **φασματική μέθοδος συσταδοποίησης** αρχικά εκτελεί φασματική ενσωμάτωση (μείωση της διαστατικότητας) στα αρχικά δεδομένα και στη συνέχεια εφαρμόζει ένα παραδοσιακό αλγόριθμο συσταδοποίησης (για παράδειγμα τον αλγόριθμο K-means) στον διαστατικά μειωμένο χώρο. Πρόσφατες έρευνες που έγιναν πάνω στη φασματική συσταδοποίηση, ανέδειξαν την ικανότητα χειρισμού μεγάλου βαθμού μη γραμμικών δεδομένων (ο χώρος δεδομένων έχει υψηλή καμπυλότητας σε κάθε τοπική περιοχή). Οι ισχυρές διασυνδέσεις της μεθόδου με την διαφορική γεωμετρία, την καθιστούν ικανή να ανακαλύψει πολλαπλές δομές του χώρου των εγγράφων. Ένα σημαντικό μειονέκτημα όμως στους αλγόριθμους φασματικής συσταδοποίησης είναι η χρήση μη γραμμικής ενσωμάτωσης (μείωσης της διαστατικότητας) που ορίζεται μόνο σε δοκιμαστικά δεδομένα. Κατά συνέπεια πρέπει να χρησιμοποιούν όλους τους δείκτες δεδομένων, ώστε να αφομοιώσουν αυτοί οι αλγόριθμοι την ενσωμάτωση. Όταν το σύνολο δεδομένων είναι πολύ μεγάλο, είναι υπολογιστικά ασύμφορο να αφομοιωθεί η ενσωμάτωση, με αποτέλεσμα να περιορίζεται η εφαρμογή της μεθόδου σε μεγάλα σύνολα.

Το **μίγμα μοντέλων συσταδοποίησης** προτυποποιεί τα δεδομένα με συνδυασμό μοντέλων και περιλαμβάνει συνήθως μοντέλα πολυωνυμικών παραγόντων. Η συσταδοποίηση περιλαμβάνει δυο βήματα: (1) εκτίμηση των παραμέτρων του μοντέλου, βασισμένη στα δεδομένα των κειμένων και σε οποιαδήποτε προηγούμενη γνώση και, (2) συναγωγή των συστάδων που βασίζονται στις εκτιμώμενες παραμέτρους. Δεδομένου το πώς έχει οριστεί το μίγμα μοντέλων οι μέθοδοι αυτές μπορούν να συσταδοποιήσουν λέξεις και έγγραφα ταυτόχρονα. Η ***Πιθανοτική Λανθάνουσα Σημασιολογική Δεικτοδότηση*** *(**PLSA**)* και η ***Λανθάνουσα Κατανομή Dirichlet*** *(Latend Dirichlet Association)* είναι δυο παραδείγματα τέτοιων μεθόδων. *'*Ένα πιθανό πλεονέκτημα των παραπάνω μεθόδων συσταδοποίησης είναι ότι οι συστάδες μπορούν να σχεδιαστούν με τέτοιο τρόπο ώστε να διευκολυνθεί η συγκριτική ανάλυση των εγγράφων.

Οι μέθοδοι **LSI** και **LPI**, όπως έχουν περιγραφεί στις ενότητες 1.3.6.1 και 1.3.6.2 αντίστοιχα, αποτελούν γραμμικές μεθόδους μείωσης της διαστατικότητας. Με αυτές τις μεθόδους μπορούμε να χρησιμοποιήσουμε τα διανύσματα μετασχηματισμού (συναρτήσεις ενσωμάτωσης), ώστε να «εκπαιδευτεί» μέρος των δεδομένων στην ενσωμάτωση και να ενσωματωθούν όλα τα δεδομένα σε χαμηλότερο διαστατικά χώρο. Με αυτό το τέχνασμα η **συσταδοποίηση με χρήση των μεθόδων LSI και LPI** μπορεί να χειριστεί μεγάλο όγκο δεδομένων από έγγραφα. Όπως αναφέρθηκε και προηγουμένως, η μέθοδος **LSI** αναζητά τα πιο χαρακτηριστικά στοιχεία σε αντίθεση με την μέθοδο **LPI** που στοχεύει στον εντοπισμό της τοπικής γεωμετρικής δομής του χώρου που ορίζουν τα έγγραφα, με αποτέλεσμα να έχει περισσότερη διακριτική ικανότητα σε σχέση με την **LSI**. Για τους σκοπούς της συσταδοποίησης, η **LPI** σαν μέθοδος μείωσης της διαστατικότητας είναι πιο κατάλληλη από την **LSI**. Σε σύγκριση όμως και με τις δυο η **PLSI** αποκαλύπτει τις λανθάνουσες σημασιολογικές διαστάσεις με πιο κατανοητό τρόπο και μπορεί εύκολα να επεκταθεί ώστε να ενσωματώσει οποιεσδήποτε προηγούμενες γνώσεις ή προτιμήσεις σχετικά με τη συσταδοποίηση.

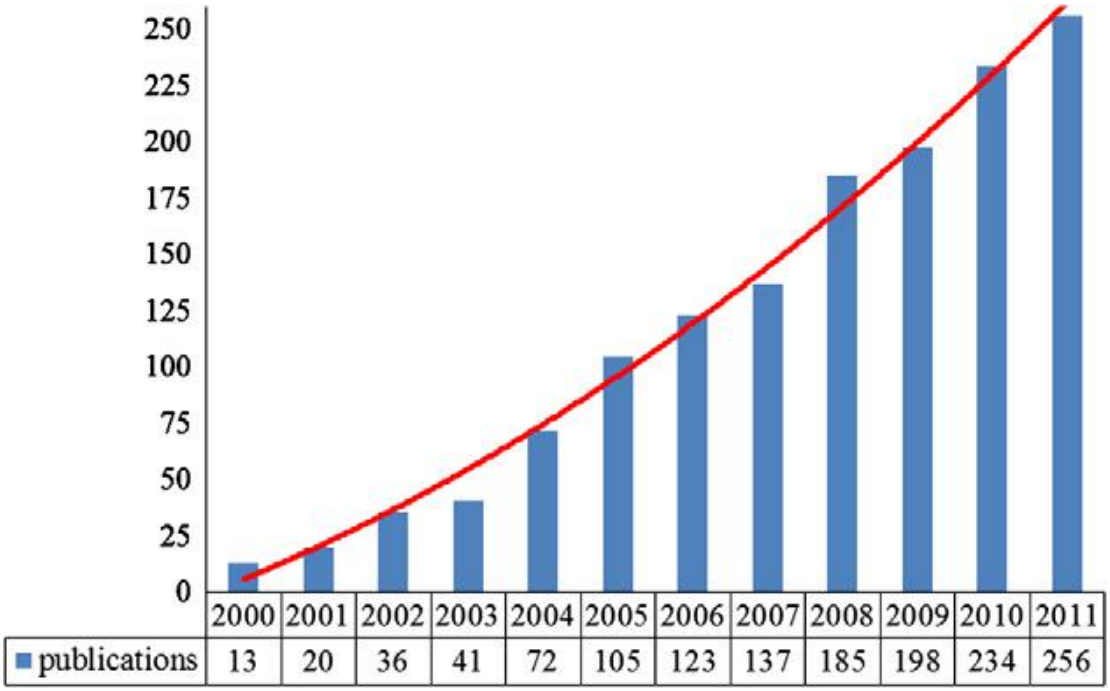
Κεφάλαιο 2: Εξόρυξη Γνώσης από Κείμενα – Εφαρμογές στην Βιοιατρική

2.1 Εισαγωγή

Η Εξόρυξη Γνώσης από Κείμενα καλύπτει ένα ευρύ φάσμα δραστηριοτήτων και μια συστοιχία διαδικασιών, αλλά ουσιαστικά στόχος της είναι να βοηθήσει τους χρήστες να ασχοληθούν με την υπερφόρτωση και παράβλεψη πληροφοριών[5]. Βασικές πτυχές της είναι να ανακαλύψει ανυποψίαστες, νέες γνώσεις κρυμμένες στην αχανή επιστημονική βιβλιογραφία, για να υποστηρίξει την ανακάλυψη υποθέσεων με γνώμονα τα δεδομένα και να αντλήσει νόημα από την πλούσια γλώσσα των ειδικών ιατρών, όπως εκφράζεται σε πληθώρα έγγραφων εκθέσεων, άρθρων, ερευνών και άλλων κειμένων. Με περίπου το 80% των έγγραφων πληροφοριών σε αδόμητη μορφή και τον ολοένα αυξανόμενο αριθμό δημοσιεύσεων (εκτιμάται ότι περίπου 2.5 εκατομμύρια άρθρα δημοσιεύονται ετησίως[6]), δεν πρέπει να προκαλεί έκπληξη το γεγονός ότι πολύτιμες νέες πηγές ερευνητικών δεδομένων παραμένουν τυπικά ανεκμετάλλευτες και ψήγματα διορατικότητας ή νέας γνώσης δεν ανακαλύπτονται σχεδόν ποτέ. Οι επιστήμονες δεν είναι σε θέση να παρακολουθούν τις εξελίξεις στα πεδία τους και να κάνουν τις συνδέσεις μεταξύ των φαινομενικά άσχετων γεγονότων για την παραγωγή νέων ιδεών και υποθέσεων.

Οι τεράστιοι αριθμοί βιοιατρικών κειμένων παρέχουν μια πλούσια πηγή γνώσης για τη βιοιατρική έρευνα. Η εξόρυξη γνώσης από κείμενα προσφέρει μια λύση για να λυθεί αυτό το πρόβλημα με την αντικατάσταση ή τη συμπλήρωση του ανθρώπινου παράγοντα, με αυτοματοποιημένο τρόπο ώστε να μετατρέψει τα αδόμητα κείμενα και την κρυμμένη γνώση, σε δομημένα δεδομένα και κατ' αυτόν τον τρόπο σε ρητή γνώση[7]. Όπως φαίνεται στην **Εικόνα 4**, ο αριθμός των δημοσιεύσεων που προέκυψαν από τη χρήση του PubMed[8]©, χρησιμοποιώντας ως ερώτημα τη λέξη κλειδί "text mining" στον τίτλο ή την περίληψη ενός κειμένου, έχει αυξηθεί ουσιαστικά από το 2000. Πολλοί ερευνητές έχουν επωφεληθεί από την τεχνολογία εξόρυξης κειμένων για να ανακαλύψουν νέα γνώση προς βελτίωση των ιατροβιολογικών ερευνών, ιδίως εκείνων που αφορούν την ανάπτυξη κακοήθων νόσων, όπως ο καρκίνος[9].

Στην επόμενη ενότητα, θα παρουσιάσουμε τις θεμελιώδεις έννοιες και διαδικασίες της εξόρυξης γνώσης από κείμενα. Θα παρουσιαστούν μερικοί αντιπροσωπευτικοί αλγόριθμοι για κάθε σημαντική διαδικασία και θα εξεταστεί κατά πόσο οι αλγόριθμοι αυτοί έχουν χρησιμοποιηθεί στην εξόρυξη γνώσης από βιοιατρικά κείμενα. Στη συνέχεια θα παρουσιαστούν κάποιες state-of-the-art εφαρμογές και σύνολα δεδομένων, ειδικά εκείνα που αναπτύχθηκαν για την γονιδιωματική και μετα-γονιδιωματική εποχή.



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
publications	13	20	36	41	72	105	123	137	185	198	234	256

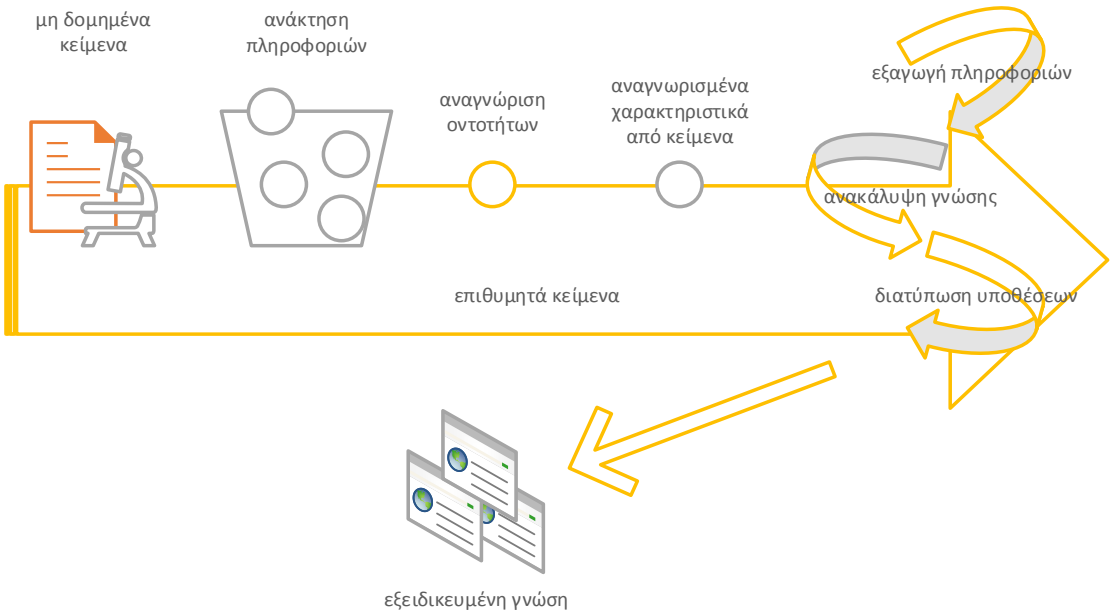
Εικόνα 4 Αριθμός Δημοσιεύσεων στο PubMed©, χρησιμοποιώντας τη λέξη-κλειδί "text mining" στον τίτλο ή την περίληψη

2.2 Στάδια και Διαδικασίες στην Εξαγωγή Γνώσης από Βιοιατρικά Κείμενα

Ο στόχος της εξόρυξης ενός κειμένου είναι να αντλήσει τη σιωπηρή γνώση που κρύβεται σε μη δομημένα κείμενα και να τα παρουσιάσει σε ρητή μορφή. Αυτό γενικά περιλαμβάνει τα παρακάτω τέσσερα στάδια:

- **ανάκτηση πληροφοριών (Information Retrieval)**,
- **εξαγωγή πληροφοριών (Information Extraction)**,
- **ανακάλυψη γνώσης (Knowledge Discovery)**, και
- **διατύπωση υποθέσεων (Hypotheses Generation).**

Τα συστήματα ανάκτησης πληροφοριών στοχεύουν στο να συλλέξουν τα επιθυμητά κείμενα για ένα συγκεκριμένο θέμα, ενώ τα συστήματα εξαγωγής πληροφοριών χρησιμοποιούνται για την εξαγωγή προκαθορισμένων τύπων πληροφοριών, όπως εξαγωγή σχέσεων. Τα συστήματα ανακάλυψης γνώσης βοηθάνε στην ανακάλυψη και εξαγωγή καινούριας γνώσης από τα κείμενα και τα συστήματα διατύπωσης υποθέσεων στην εξαγωγή συμπερασμάτων για άγνωστα βιοιατρικά γεγονότα που βασίζονται σε κείμενο, όπως φαίνεται στην **Εικόνα 5**. Ουσιαστικά δηλαδή, οι γενικές αποστολές της εξόρυξης γνώσης από βιοιατρικά κείμενα, περιλαμβάνουν την ανάκτηση πληροφοριών, την αναγνώριση ονοματικών οντοτήτων και εξαγωγή σχέσεων, την ανακάλυψη γνώσης και την διατύπωση υποθέσεων.



Εικόνα 5 Συνηθισμένα Στάδια και Διαδικασίες στην εξόρυξη γνώσης από βιοιατρικά κείμενα

2.2.1 Ανάκτηση Πληροφοριών

Εκτός από τα συμβατικά συστήματα ανάκτησης πληροφοριών, υπάρχουν επίσης συστήματα ανάκτησης πληροφοριών προηγμένης γνώσης, που ενσωματώνουν δεδομένα από διάφορους πόρους σε ένα ενιαίο πλαίσιο για την ενίσχυση της κατανόησης σύνθετων βιοιατρικών συστημάτων. Για παράδειγμα, για να έχει πρόσβαση σε αποτελέσματα από εξόρυξη γνώσης κειμένων αλλά και άλλα δεδομένα, δημιουργήθηκε μια βάση γνώσης[10] για χρόνια αποφρακτική πνευμονοπάθεια και αναπτύχθηκε κατ' αυτό τον τρόπο ένα ολοκληρωμένο σύστημα διαχείρισης γνώσης. Η γνωσιακή βάση **Salivaomics©**[11] (salivaomics ή salivary diagnostics είναι μια διαδικασία για εξέταση και διατήρηση δεδομένων για διαγνωστικούς σκοπούς που στηρίζεται σε αισθητήρες, οι οποίοι έχουν ως βάση το σίελο) ορίζει την οντολογία «*Saliva*» ως ένα λεξιλόγιο όρων και σχέσεων για τη διευκόλυνση της ανάκτησης δεδομένων και της ολοκλήρωσης τους σε πολλαπλά πεδία της έρευνας καθώς και στην ανάλυση και εξόρυξη δεδομένων. Το **QuExT**[12], ένα σύστημα ανάκτησης εγγράφων που βασίζεται στο **PubMed©**, ακολούθησε μια μεθοδολογία με επέκταση σε ερωτήματα προσανατολισμένα στην έννοια, για να εντοπίζει έγγραφα που περιέχουν έννοιες που σχετίζονται με τις λέξεις-κλειδιά. Στη γονιδιωματική εποχή, με τις τελευταίες προόδους στον τομέα της βιοτεχνολογίας και των μεθοδολογιών υψηλού ρυθμού απόδοσης για την ανάλυση του γονιδίου, θα υπάρξει μια συνεχώς αυξανόμενη ανάγκη για εργαλεία εξόρυξης γνώσης από κείμενα και ανάκτησης πληροφοριών, ώστε να βοηθήσει τους ερευνητές να βρουν άρθρα σχετικά με τις μελέτες τους.

2.2.2 Αναγνώριση Ονοματικών Οντοτήτων και Εξαγωγή Σχέσεων

Η **Αναγνώριση Ονοματικών Οντοτήτων (Named Entity Recognition)** είναι το πιο σημαντικό στάδιο στην εξόρυξη γνώσης και έχει ως συνολικό στόχο την αναγνώριση συγκεκριμένων όρων, όπως γονίδιο, πρωτεΐνες, ασθένεια και φάρμακα. Αρκετές τεχνολογίες στην επιστήμη των υπολογιστών έχουν ασχοληθεί με την αναγνώριση βιοιατρικών όρων. Ωστόσο, στην πράξη εξακολουθούν να υπάρχουν πολλά εμπόδια για την αυτόματη αναγνώριση βιοιατρικών όρων. Για παράδειγμα, ένας όρος της βιοιατρικής μπορεί να έχει πολλές διαφορετικές γραπτές μορφές, όπως για παράδειγμα το ότι οι όροι «επιληψία» και «ασθένεια της πτώσης» αναφέρονται στην ίδια ασθένεια, η οποία είναι μια διαταραχή του κεντρικού νευρικού συστήματος που χαρακτηρίζεται από απώλεια της συνείδησης και σπασμούς. Επιπλέον, μια οντότητα μπορεί να έχει πολλαπλές σημασίες, για παράδειγμα ο καρκίνος μπορεί να αναπαρασταθεί ως μια ασθένεια, αλλά και ως ένα αστρονομικό σύμβολο. Ακόμα και συντομογραφίες των όρων μπορεί να προκαλέσουν προβλήματα ασάφειας, όπως για παράδειγμα, το PC μπορεί να σημαίνει τον καρκίνο του προστάτη (Prostate Cancer), φωσφατιδυλική χολίνη (Phosphatidyl Choline), ή ακόμη και τον προσωπικό υπολογιστή (Personal Computer). Πολλοί βιοιατρικοί όροι αποτελούνται επίσης από φράσεις ή σύνθετες λέξεις ή μπορεί να έχουν ένα πρόσφυμα.

Οι τρέχουσες τεχνικές αναγνώρισης βιοιατρικών ονοματικών οντοτήτων εμπίπτουν σε τρεις μεγάλες κατηγορίες:

- **προσεγγίσεις που βασίζονται σε λεξικό,**
- **προσεγγίσεις που βασίζονται σε κανόνες**, και
- **προσεγγίσεις μηχανικής μάθησης**.

Οι **προσεγγίσεις που βασίζονται σε λεξικό** τείνουν να χάνουν απροσδιόριστους όρους που δεν αναφέρονται στο λεξικό, ενώ οι **προσεγγίσεις που βασίζονται σε κανόνες** απαιτούν κανόνες που να αναγνωρίζουν όρους από το κείμενο, αν και οι κανόνες που προκύπτουν δεν είναι συχνά αποτελεσματικοί σε όλες τις περιπτώσεις. Οι **προσεγγίσεις μηχανικής μάθησης** γενικά απαιτούν τυποποιημένα επεξηγηματικά δοκιμαστικά σύνολα δεδομένων τα οποία συνήθως απαιτούν τεράστιες ανθρώπινες προσπάθειες για να κατασκευαστούν. Επιπλέον, οι περισσότερες προσεγγίσεις μηχανικής μάθησης τείνουν να βασίζονται σε δεδομένα και εφαρμογές προσανατολισμένα σε κάποιο συγκεκριμένο τομέα με τα μεγέθη της **ακρίβειας**, του **ρυθμού ανάκλησης** και του $\boldsymbol{F_{score}}$ που χρησιμοποιούνται συχνά για να αξιολογήσουν την απόδοση της αναγνώρισης, όπως είχαν οριστεί στην ενότητα 1.3.2, αλλά αυτή τη φορά αναφερόμενα ως προς τους όρους αναζήτησης:

$$ακρίβεια = \frac{|\{Σχετικά\} \cap \{Ανακτημένα\}|}{|\{Ανακτημένα\}|}$$

$$ανάκληση = \frac{|\{Σχετικά\} \cap \{Ανακτημένα\}|}{|\{Σχετικά\}|}$$

$$F_{score} = \frac{ακρίβεια \times ανάκληση}{(ακρίβεια + ανάκληση)/2}$$

Για παράδειγμα, όταν αναγνωριστεί ένας όρος γονιδίου, οι σχετικοί όροι είναι αυτοί οι όροι που σωστά έχουν αναγνωρισθεί ως γονίδιο, ενώ οι ανακτημένοι όροι είναι όλοι οι όροι που, ανεξάρτητα σωστά ή λάθος, έχουν αναγνωρισθεί ως γονίδιο. Οι προσεγγίσεις μηχανικής μάθησης χρησιμοποιούνται ευρέως τα τελευταία χρόνια για την αναγνώριση ονοματικών οντοτήτων, όπως για παράδειγμα τα **Κρυφά Μαρκοβιανά Μοντέλα (Hidden Markov Models, HMM)**, **Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines, SVMs)**, **Τυχαία Πεδία Υπό Συνθήκη (Conditional Random Fields, CRFs)** και συστήματα **Μέγιστης Εντροπείας (Maximum Entropy, ME)**. Επί του παρόντος, οι καλύτερες τιμές του δείκτη F_{score} για τα συστήματα αναγνώρισης βιοιατρικών ονοματικών οντοτήτων δεν είναι τόσο καλές όσο τα αποτελέσματα από τα συστήματα γενικής χρήσης. Οι ερευνητές έχουν δοκιμάσει πολλές μεθόδους για να βελτιωθεί η απόδοση, συνδυάζοντας διαφορετικές προσεγγίσεις και προτάσεις υβριδικών προσεγγίσεων, τη διεξαγωγή περαιτέρω επεξεργασίας μετά από τη μηχανική μάθηση και την προσθήκη βασικών βιοιατρικών γνώσεων.

Ένας βιοιατρικός όρος μπορεί να εμφανιστεί με τη μορφή συντομογραφίας και μπορεί επίσης να έχει πολλαπλά συνώνυμα μέσα σε ένα κείμενο. Η αναγνώριση συντομογραφιών και συνωνύμων είναι χρήσιμη για την ενοποίηση και κανονικοποίηση βιοιατρικών όρων στην αναγνώριση ονοματικών οντοτήτων. Υπάρχουν πολλά τέτοια συστήματα, όπως για παράδειγμα το **BIOADI**[13] , ένα σύστημα αναγνώρισης βιοιατρικών συντομογραφιών με βάση μια προσέγγιση μηχανικής μάθησης και το **ALICE**[14], ένα σύστημα που εξάγει ζεύγη συντομογραφιών χρησιμοποιώντας πρότυπα και κανόνες.

Οι πιο πρόσφατες έρευνες προσανατολίζονται πλέον σε αναγνώριση όρων και κανονικοποίηση και αναδύονται συνεχώς νέα συστήματα, όπως το **GeneTUKit**[15] και το **BioCreative III**: ένα από τα καθήκοντα του **BioCreative III**[16], επικεντρώνεται στην εξομάλυνση των γονιδίων, προσδιορίζει αναφορές του γονιδίου και συνδέει τα γονίδια αυτά σε πρότυπα αναγνωριστικά (για παράδειγμα, σε μια βάση δεδομένων αναγνωριστικών).

Τα συμβατικά συστήματα εξαγωγής σχέσεων επικεντρώνονται στη διερεύνηση εξαγωγής βιοιατρικών σχέσεων (για παράδειγμα αλληλουχία και αλληλεπίδραση πρωτεϊνών και σχέσεις γονιδίων με ασθένειες) από βιοιατρικούς όρους (γονίδια, πρωτεΐνες, ασθένειες ή φάρμακα). Εκτός όμως από την απλή αναγνώριση ύπαρξης των σχέσεων, ιδιαίτερα σημαντική είναι και η κατηγοριοποίηση του είδους της σχέσης, καθώς και το ότι με την βελτίωση της απόδοσης αναγνώρισης όρων μπορεί να βελτιωθεί η ακρίβεια των εξαγόμενων σχέσεων. Ως παράδειγμα, το σύστημα **MeTAE** (Medical Texts Annotation and Exploration)[17], είναι σε θέση να προσδιορίζει τη σωστή σημασιολογική σχέση μεταξύ κάθε ζεύγους οντοτήτων, χρησιμοποιώντας το **MetaMap**[18] για τον προσδιορισμό φαρμακευτικών ουσιών, ενώ μια προσέγγιση γλωσσικών προτύπων καθορίζει τη σημασιολογική σχέση μεταξύ κάθε ζεύγους.

Στη σημερινή γονιδιακή εποχή, πολλοί ερευνητές ενδιαφέρονται για εξόρυξη αλληλεπιδράσεων μεταξύ των γονιδίων, των πρωτεϊνών και άλλων ευρέων γονιδιακών σχέσεων, που παρέχουν χρήσιμες πληροφορίες για περαιτέρω, πιο περιεκτική ανάλυση της λειτουργίας των γονιδίων και χρησιμοποιούνται επεξηγηματικά σε βάσεις δεδομένων γονιδίων καθώς και σε άλλες εκτεταμένες σχέσεις. Στον τομέα αυτό έχουν εφαρμοστεί

συνδυαστικές τεχνικές εξόρυξης γνώσης από κείμενα με ανάλυση ακολουθιών για την ανακάλυψη υπομοριακών πρωτεϊνικών περιοχών[19], αλλά και η δημιουργία ενός πλαισίου εξόρυξης κειμένων και οπτικοποίησης των αποτελεσμάτων για να εντοπίσει τις λεπτομέρειες της αλληλεπίδρασης μεταξύ των πρωτεϊνών, έτσι ώστε να παρέχει μια βαθύτερη κατανόηση της πρωτεϊνικής λειτουργίας, προσδιορίζοντας την ακολουθία των αμινοξέων στην διεπαφή της αλληλεπίδρασης μιας πρωτεΐνης[20]. Επιπλέον, έχουν αναπτυχθεί συστήματα που μπορούν να χρησιμοποιηθούν ώστε να καθοριστεί εάν ένα άρθρο σχετίζεται με πρωτεϊνικές αλληλεπιδράσεις και εφόσον σχετίζεται, να αντιστοιχιστεί αυτή η σχέση με άλλα σχετικά άρθρα[21].

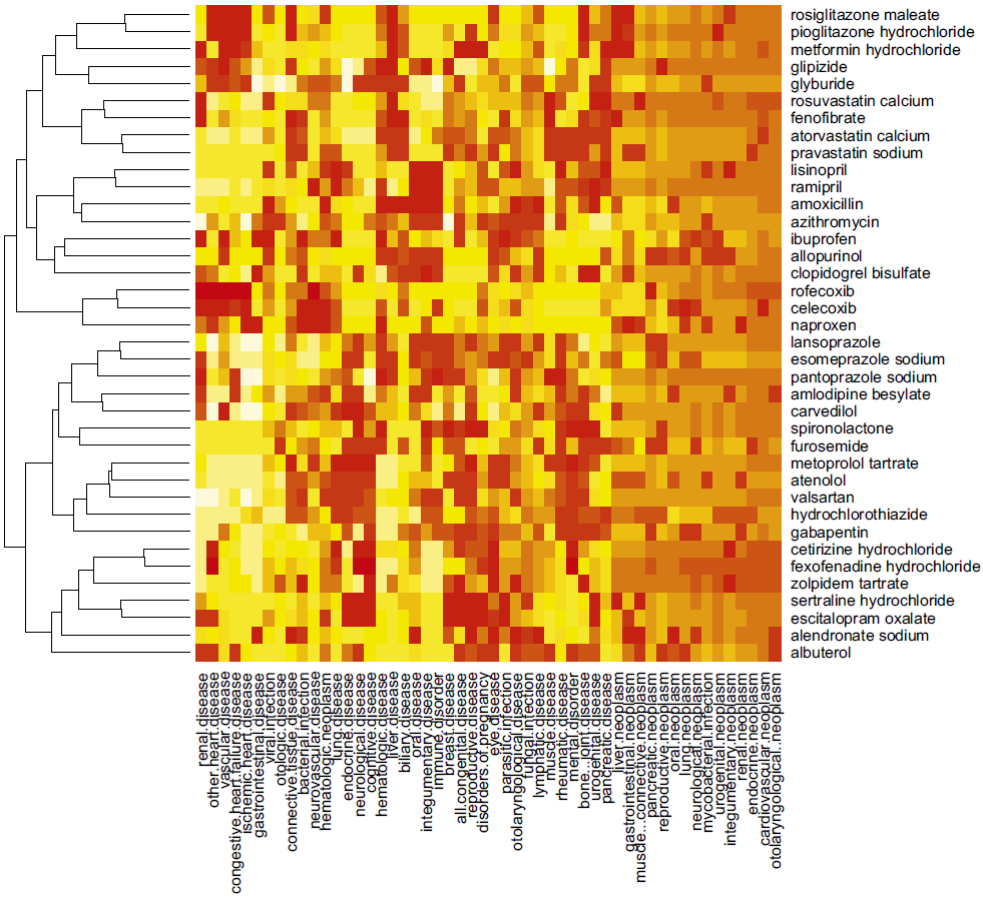
Παράλληλα, οι ερευνητές επικεντρώνονται στη σχέση μεταξύ γονιδίων και πρωτεϊνών και άλλων οντοτήτων της βιοιατρικής, όπως σχέσεις γονιδίων με ασθένειες και υποκυτταρικές σχέσεις πρωτεϊνών, παραδείγματος χάριν το διαδικτυακό σύστημα **PLAN2L©**[22], με ενσωματωμένη εξόρυξη γνώσης από κείμενα και εξαγωγή σχέσεων μεταξύ βιοιατρικών οντοτήτων που προέρχονται από τη λογοτεχνία και έχει συστηματικά πρόσβαση σε πληροφορίες για την ανάλυση της γενετικών, κυτταρικών, και μοριακών πτυχών του φυτικού οργανισμού *Arabidopsis thaliana*. Η έρευνα των **Shetty KD** και **Dalal SR** «**Using information mining of the medical literature to improve drug safety, J Am Med Inform Assoc 2011; 18:668–74**» απέδειξε ότι η ανάλυση της ιατρικής λογοτεχνίας σε ένα πλαίσιο δυσαναλογίας μπορεί να υποστηρίξει σύγχρονες μεθόδους για ανακάλυψη σχέσεων μεταξύ φαρμάκων και παρενεργειών, συμβάλλοντας κατ' αυτόν τον τρόπο στη βελτίωση της πολιτικής ασφαλείας για τα φάρμακα. Τα αποτελέσματα αυτής της έρευνας φαίνονται στην **Εικόνα 6**, με την μορφή θερμικού χάρτη. Κάθε κελί του πίνακα αλληλοσυσχέτισης παρενεργειών και φαρμάκων αντιπροσωπεύει την πιθανότητα της παρατηρούμενης αρίθμησης να υπερβαίνει τον αναμενόμενο αριθμό, υποθέτοντας την ανεξαρτησία των δυσμενών αποτελεσμάτων - παρενεργειών και των φαρμάκων. Οι τιμές που εμφανίζεται ένα φάρμακο αναπαριστώνται χρησιμοποιώντας ένα φάσμα χρωμάτων, από χαμηλές (κόκκινο) σε υψηλές (κίτρινο). Τα φάρμακα συσταδοποιήθηκαν βάσει της ομοιότητας τους σε παρενέργειες, ενώ οι παρενέργειες συσταδοποιήθηκαν βάσει της ομοιότητας τους σε φάρμακα.

2.2.3 Ανακάλυψη Γνώσης

Η γνώση, συμπεριλαμβανομένων γεγονότων, πληροφοριών ή περιγραφών, σιωπηρή ή ρητή, αναφέρεται στην θεωρητική ή πρακτική κατανόηση ενός τομέα ή ενός θέματος. Η ανακάλυψη της γνώσης είναι η δημιουργία γνώσης από μεγάλους όγκους δομημένων ή μη δομημένων δεδομένων. Η γνώση που αποκτήθηκε μπορεί να μετατραπεί σε συμπληρωματικά δεδομένα, που μπορούν να χρησιμοποιηθούν για περαιτέρω επεξεργασία ή ανακάλυψη. Η ανακάλυψη γνώσης είναι ένα πολύ σημαντικό μέρος της εξόρυξης δεδομένων. Πιο συγκεκριμένα, η ανακάλυψη γνώσης από κείμενα βιοιατρικής είναι μια διαδικασία με στόχο να βρει απαντήσεις για ερωτήσεις της βιοιατρικής, όπως ο προσδιορισμός στόχων νέων φαρμάκων ή καινοτόμους βιοδείκτες διάγνωσης καρκίνου. Το **CRAB**[23], ένα πλήρως ολοκληρωμένο εργαλείο εξόρυξης κειμένων, εξάγει τα σχετικά στοιχεία στη λογοτεχνία για τον κίνδυνο καρκίνου και τα αξιολογεί χρησιμοποιώντας τεχνικές ανακάλυψης γνώσης. Με τη χρήση του **CRAB** αποδείχτηκε ότι η χρήση τεχνικών διοχέτευσης (*pipeline*) στην εξόρυξη κειμένων μπορεί να διευκολύνει πολύπλοκες ερευνητικές εργασίες στη βιοιατρική. Η εξόρυξη γνώσης από κείμενα βοήθησε στην ανακάλυψη δύο τρόπων για την λειτουργική συμμετοχή ενός συνόλου γονιδίων πρόβλεψης, που φανερώνουν

επιδεκτικότητα για πρώιμη εμφάνιση καρκίνου του παχέος εντέρου, ξεπερνώντας την έλλειψη σε μελέτες έκφρασης ολόκληρου του γονιδιώματος για τον καρκίνο του παχέος εντέρου[24].

Η ανακάλυψη γνώσης είναι σε θέση να ενσωματώσει βιοιατρικά κείμενα με πολλαπλές πηγές δεδομένων και να δημιουργήσει ένα καινοτόμο ερμηνευτικό πλαίσιο. Για παράδειγμα, μέσα από την τεχνολογία εξόρυξης κειμένου σε συνδυασμό με δεδομένα μικροσυστοιχιών, ανακαλύφθηκε ότι μετα-μεταγραφικοί έλεγχοι διαδικασιών των ωοθηκών είναι πιθανή αιτία για παρατηρούμενους ογκολογικούς και αναπαραγωγικούς φαινότυπους, όπως και το ότι οι επαναλαμβανόμενοι έμμηνοι κύκλοι αποτελούν τον πραγματικό σύνδεσμο μεταξύ ογκογενέσεων και αναπαραγωγικών ελέγχων[25].



Εικόνα 6 *Θερμικός Χάρτης Πίνακα Αλληλοσυσχέτισης Παρενεργειών- Φαρμάκων*

2.2.4 Διατύπωση Υποθέσεων

Με βάση τα γεγονότα ή τις πληροφορίες που δεν μπορούν να εξηγηθούν ικανοποιητικά με τις διαθέσιμες γνώσεις, μπορεί να προταθεί μια **επιστημονική υπόθεση**, που είναι μια δοκιμαστική λύση σε ένα πρόβλημα παρά μια θεωρία, με πρόταση για περαιτέρω έρευνα. Αποτελέσματα και διεξαγωγές πειραμάτων μπορούν να χρησιμοποιηθούν για την αξιολόγηση της προτεινόμενης υπόθεσης πριν από την επίλυση του προβλήματος. Η

επιστημονική υπόθεση είναι κατά κάποιον τρόπο σαν υλοποίηση μιας επιστημονικής φαντασίας που βασίζεται σε υπάρχοντα στοιχεία και γνώσεις. Η διατύπωση μιας υπόθεσης είναι το να πάρει μη αποδεδειγμένη συμπερασματολογία από στοιχεία κρυμμένα στο κείμενο, ενώ η ανακάλυψη γνώσης είναι η εξαγωγή καινούριας γνώσης.

Η βιοιατρική βιβλιογραφία είναι μια αστείρευτη πηγή πιθανών πληροφοριών για την εξαγωγή συμπερασμάτων και την παραγωγή νέων υποθέσεων. Η διατύπωση υποθέσεων είναι μια σημαντική λειτουργία στην εξόρυξη κειμένων, που είναι πολύ χρήσιμη για βιοιατρικούς ερευνητές, οι οποίοι θέλουν να βγάλουν συμπεράσματα πάνω σε άγνωστα βιοιατρικά γεγονότα, που μπορούν να χρησιμοποιηθούν ώστε να καθοδηγήσουν το σχεδιασμό επιπλέον πειραμάτων ή να εξηγήσουν τα υπάρχοντα πειραματικά αποτελέσματα. Αυτή η εργασία απολαμβάνει σταδιακά μεγαλύτερη προσοχή από τους ερευνητές. Για παράδειγμα, μέσα από την εξερεύνηση χαρτών συνδεσιμότητας φαρμάκων-πρωτεϊνών, ειδικά για την νόσο του Alzheimer, με βάση δίκτυα αλληλεπίδρασης πρωτεϊνών και εξόρυξη κειμένων, προέκυψε μια νέα υπόθεση ότι μπορεί να διερευνηθεί η διλτιαζέμη (diltiazem) και η κινιδίνη (quinidine) ως υποψήφια φάρμακα για τη θεραπεία της νόσου του Alzheimer[26]. Αποτέλεσμα μιας άλλης έρευνας ήταν να δημιουργηθούν δίκτυα πρόβλεψης αλληλεπίδρασης διαφορετικών ειδών καρκίνου, μέσα από το συνδυασμό αποτελεσμάτων από εξόρυξη γνώσης από βιοιατρικά κείμενα και από δεδομένα για δομικές αλληλεπιδράσεις πρωτεϊνών[27].

2.3 Σύνολα Δεδομένων και Εργαλεία για Εξόρυξη Γνώσης από Βιοιατρικά Κείμενα

Όσον αφορά τα συστήματα ανάκτησης πληροφοριών, το **PubMed©** είναι μια από τις πιο γνωστές βιοιατρικές βάσεις δεδομένων και περιέχει περισσότερα από 20 εκατομμύρια παραπομπές για βιοιατρικά άρθρα από το **MEDLINE®**[28] και διάφορα περιοδικά βιολογικής επιστήμης, η οποία παρέχει μια εύχρηστη διαδικτυακή πύλη αναζήτησης για τους χρήστες, καθώς και μια διεπαφή προγράμματος εφαρμογής για τους προγραμματιστές. Το **Textpresso**[29] χρησιμοποιεί μια οντολογία και επιστρέφει τα αποτελέσματα αναζήτησης για τις κατηγορίες των βιολογικών εννοιών (π.χ. γονίδιο, αλληλόμορφο κύτταρο, φαινότυπος), τις τάξεις των σχέσεων των αντικειμένων (π.χ., ένωση, κανονισμός) και τις σχετικές περιγραφές (π.χ., βιολογική διαδικασία). Το **GoPubMed®**[30] κατατάσσει περιλήψεις από βιοιατρική λογοτεχνία, σύμφωνα με μια Οντολογία Γονιδίων και δείχνει τους όρους της οντολογίας που σχετίζονται με τις λέξεις-κλειδιά του ερωτήματος. Επιπλέον, επιτρέπει στους χρήστες να εξερευνήσουν αποτελέσματα αναζήτησης στο **PubMed©** με ένα πρόγραμμα προβολής οντολογιών.

Για την αναγνώριση των βιοιατρικών οντοτήτων, υπάρχουν αρκετά ισχυρά συστήματα και σύνολα δεδομένων. Στον **Πίνακα 1** περιέχεται μια λίστα με μερικά χρήσιμα συστήματα αναγνώρισης βιοιατρικών οντοτήτων. Ο **Πίνακας 2** περιέχει τα πρότυπα σύνολα δεδομένων που μπορούν να χρησιμοποιηθούν για να αξιολογηθεί η απόδοση ενός συστήματος αναγνώρισης ονοματικών οντοτήτων ή για να αναπτυχθεί ένα σύστημα με βάση μια μηχανή μάθησης. Οι **Πίνακες 1** και **2** επίσης, έχουν ορισμένους πόρους και συστήματα αναγνώρισης συνωνύμων και συντομογραφιών.

Υπάρχουν πολλά χρήσιμα συστήματα εξαγωγής σχέσεων, όπως το **iHOP**[31] , το οποίο ανιχνεύει τις αλληλεπιδράσεις μεταξύ των γονιδίων, χρησιμοποιώντας τα γονίδια ή τις

πρωτεΐνες ως υπερ-συνδέσεις μεταξύ προτάσεων και αποσπασμάτων με βάση μια προσέγγιση συν-εμφάνισης. Περισσότερα συστήματα εξαγωγής σχέσεων παρουσιάζονται στον **Πίνακα 3**.

Για να ξεπεραστεί η έλλειψη ολοκλήρωσης μεταξύ γονιδιακών δεδομένων και βιολογικής λογοτεχνίας, αναπτύχθηκε ένα εργαλείο που συνδέει πάνω από 2 εκατομμύρια άρθρα στο **PubMed©** με σχεδόν 150.000 γονίδια από 50 είδη στο **Ensembl**[32]. Τα σύνολα δεδομένων για την εξαγωγή σχέσεων είναι επίσης σημαντικά και μερικά από αυτά παρουσιάζονται στον **Πίνακα 4**. Τέλος, μερικά από τα συνήθως χρησιμοποιούμενα τυποποιημένα, επεξηγηματικά σύνολα δεδομένων για τους σκοπούς της εξόρυξης κειμένων παρατίθενται στον **Πίνακα 5**.

Πολλά συστήματα διατύπωσης υποθέσεων είναι διαθέσιμα. Το **BioText-Quest**[33] είναι ένα σύστημα εξόρυξης βιοιατρικών κειμένων για την ανακάλυψη εννοιών, που παρέχει υπηρεσίες όπως αναγνώριση βιοιατρικών οντοτήτων, σχέσεις εννοιών και διατύπωση υποθέσεων. Το **Arrowsmith**[34] προσδιορίζει νοηματικές συνδέσεις μεταξύ δύο συνόλων από άρθρα στο **MEDLINE®**, ενώ το **BITOLA**[35] μπορεί να χρησιμοποιηθεί ώστε να ανακαλύψει νέες σχέσεις μεταξύ βιοιατρικών οντοτήτων ή εννοιών, όπως η νόσος υποψηφίων γονιδίων στη βιοιατρική λογοτεχνία.

Πίνακας 1

Συνήθη Συστήματα Αναγνώρισης Βιοιατρικών Οντοτήτων

Σύστημα	**Σύντομη Περιγραφή**
ABNER[36]	Το ABNER είναι ένα εργαλείο λογισμικού για ανάλυση κειμένου της μοριακής βιολογίας. Χρησιμοποιεί προσέγγιση τυχαίων πεδίων γραμμικής αλυσίδας υπό όρους με ορθογραφικά και συμφραζόμενα χαρακτηριστικά.
GENIATagger[37]	Το GENIATagger είναι ειδικά ρυθμισμένο για βιοιατρικά κείμενα, όπως περιλήψεις από το **MEDLINE®**. Είναι ένα χρήσιμο εργαλείο προ-επεξεργασίας για εξαγωγή πληροφοριών από βιοιατρικά έγγραφα.
LingPipe[38]	Το LingPipe παρέχει τρία γενικά, εκπαιδεύσιμα “chunkers” για να πραγματοποιήσουν αναγνώριση ονοματικών οντοτήτων. Το LingPipe μπορεί να χρησιμοποιηθεί για τον εντοπισμό βιοιατρικών οντοτήτων όπως, γονίδια, οργανισμοί, κακοήθειες και χημικές ουσίες.
Yapex[39]	Το Yapex είναι ένα σύστημα αναγνώρισης ονοματικών οντοτήτων που λειτουργεί με ένα σύστημα κανόνων και χρησιμοποιεί λεξιλογική και συντακτική ανάλυση για τον προσδιορισμό ονομάτων πρωτεϊνών.

Πίνακας 2

Πρότυπα Επεξηγηματικά Σύνολα Δεδομένων για Αναγνώριση Βιοιατρικών Οντοτήτων

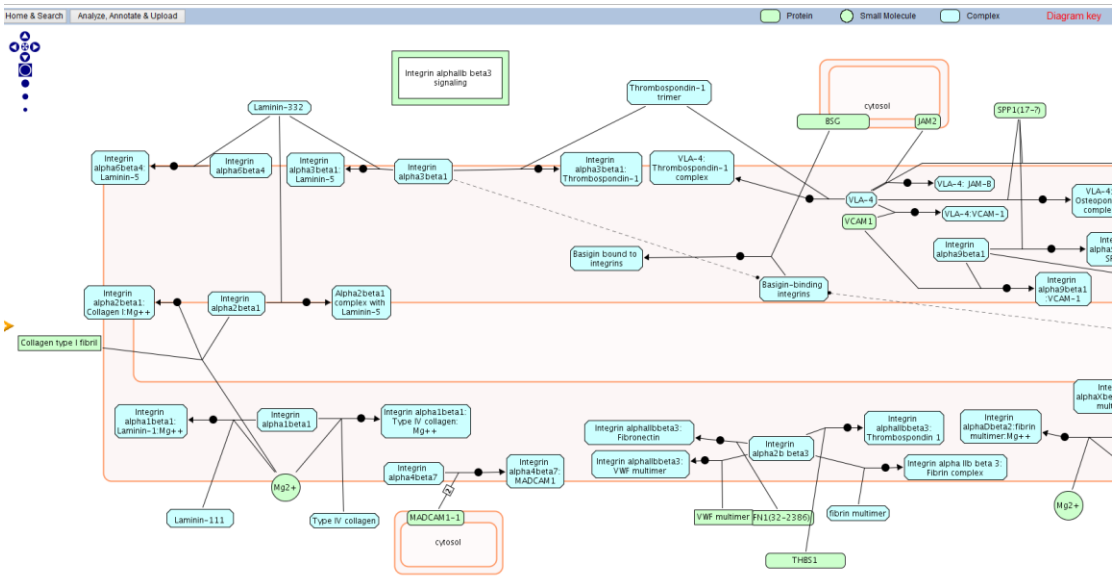
Σύστημα	**Σύντομη Περιγραφή**
Acromine[40]	Το λεξικό συντομογραφιών του Acromine κατασκευάζεται αυτόματα από ολόκληρο το MEDLINE®. Το Acromine αποδείχτηκε ότι ήταν αρκετά καλό και μετά εφαρμόστηκε σε ολόκληρο το MEDLINE®.
BioLexicon[41]	Το BioLexicon συγκεντρώνει ορολογίες από αρκετές μεγάλες δημόσιες πηγές δεδομένων βιοπληροφορικής, όπως τις UniProtKb, ChEBI και NCBI. Το BioLexicon αντιπροσωπεύει τους όρους σε συνδυασμό με λεξιλογικές και στατιστικές πληροφορίες, προκειμένου να βελτιωθεί η απόδοση της εξόρυξης κειμένων.
GENETAG[42]	Το GENETAG είναι ένα από τα σημαντικότερα τυποποιημένα πρότυπα σύνολα δεδομένων για δοκιμαστική αναγνώριση βιοιατρικών οντοτήτων. Έχει 20.000 προτάσεις από το MEDLINE® για αναγνώριση όρων γονιδίων/πρωτεϊνών.
GO[43]	Το GO (Gene Ontology) είναι μια σημαντική πρωτοβουλία της βιοπληροφορικής που αποσκοπεί στην εκπροσώπηση των γονιδίων και του γονιδιακού προϊόντος με τυποποιημένη μορφή. Το GO παρέχει ένα ελεγχόμενο λεξιλόγιο με όρους για την περιγραφή χαρακτηριστικών προϊόντος του γονιδίου και επεξηγηματικά στοιχεία των προϊόντων των γονιδίων.

Πίνακας 3

Χρήσιμα Εργαλεία για Εξαγωγή Σχέσεων

Σύστημα	**Σύντομη Περιγραφή**
BCMS[44]	Το BioCreative MetaServer (BCMS) είναι ένα meta-service για την εξαγωγή πληροφοριών που μπορεί να δημιουργήσει σχολιασμούς σε αποσπάσματα των PubMed© και Medline®, καλύπτοντας ονόματα γονιδίων, αναγνωριστικά γονιδίων, είδη και αλληλεπιδράσεις πρωτεϊνών.
Chilibot[45]	Το Chilibot εξερευνά περιλήψεις στο PubMed© για συγκεκριμένες σχέσεις μεταξύ πρωτεϊνών, γονιδίων ή λέξεων-κλειδιών. Τα αποτελέσματα επιστρέφονται ως γραφική παράσταση.
HPID[46]	Το Human Protein Interaction Database (HPID) παρέχει πληροφορίες για αλληλεπιδράσεις μεταξύ ανθρώπινων πρωτεϊνών από τα υπάρχοντα δομικά και πειραματικά δεδομένα, και ολοκληρωμένες ανθρώπινες πρωτεϊνικές αλληλεπιδράσεις που προέρχονται από τα BIND, DIP και

Σύστημα	Σύντομη Περιγραφή
	HPRD. Οι χρήστες μπορούν να βρουν πιθανές αλληλεπιδράσεις μεταξύ της πρωτεΐνης που εισάγουν σαν είσοδο, με πρωτεΐνες από τις βάσεις δεδομένων. Τα αναγνωριστικά μιας πρωτεΐνης στο EMBL, Ensembl, MIM, RefSeq, HPRD και NCBI μπορούν να χρησιμοποιηθούν κατά τη διάρκεια της αναζήτησης για αλληλεπίδραση.
HPRD[47]	Το Human Protein Reference Database (HPRD) είναι μια πλατφόρμα για δίκτυα αλληλεπίδρασης ανθρώπινων πρωτεϊνών και αλληλοσυσχέτισης με ασθένειες. Όλες οι πληροφορίες στο HRPD έχουν εισαχθεί χειροκίνητα από επιστήμονες. Για κάθε στοιχείο της πρωτεομικής το HRPD μπορεί να αναπτύξει οπτικά τα αποτελέσματα.
iHOP[48]	Το Information Hyperlinked over Proteins (iHOP) μπορεί να δημιουργήσει ένα δίκτυο συγκλίνοντων γονιδίων και πρωτεϊνών από εκατομμύρια περιλήψεις του PubMed©. Το iHOP χρησιμοποιεί τα γονίδια και τις πρωτεΐνες, ως υπερ-συνδέσεις μεταξύ προτάσεων και αποσπασμάτων, ως εκ τούτου οι πληροφορίες μπορούν να μετατραπούν σε ένα ολοκληρωμένο δίκτυο πόρων.
IntAct[49]	Το IntAct παρέχει εργαλεία ανάλυσης για μοριακή αλληλεπίδραση, καθώς και αλληλεπίδραση με βάσεις δεδομένων, των οποίων τα δεδομένα προήλθαν από επιμεληημένη λογοτεχνία ή δημοσιεύσεις χρηστών.
MedScan[50]	Στο MedScan συλλέγονται πληροφορίες και γίνεται ανάκτηση δεδομένων από πολλαπλές πηγές ενημέρωσης του κοινού, κείμενα, περιοδικά, καθώς και διάφορα σύνολα δεδομένων και στη συνέχεια μετατρέπονται σε βιολογικές σχέσεις που θα μπορούσαν να χρησιμοποιηθούν για την διατύπωση υποθέσεων και επαλήθευση, κατανόηση ασθενειών, διαχείριση ασθενών και φαρμάκων.
PubGene[51]	Τα ανακτημένα ονόματα γονιδίων και πρωτεϊνών από το PubGene διασταυρώνονται μεταξύ τους και με σχετικούς όρους, με στόχο την κατανόηση της βιολογικής λειτουργίας τους, τη σημασία τους σε μια ασθένεια και τη σχέση τους.
Reactome[52]	Το Reactome αποτελείται από εργαλεία ανάλυσης δεδομένων ανοιχτού κώδικα, καθώς και μια μη αυτόματα επιμελημένη και αναθεωρημένη βάση δεδομένων που συμπεριλαμβάνει δεδομένα αλληλεπίδρασης, αντίδρασης και «μονοπατιών» (pathway data). Το Reactome μπορεί να χρησιμοποιηθεί για την αλληλεπίδραση, την αντίδραση και την ανάλυση με βάση το μονοπάτι, όπως φαίνεται στην **Εικόνα7**.



Εικόνα 7 *Μέρος Μονοπατιού των Αλληλεπιδράσεων ενός Integrin στην Κυτταρική Επιφάνεια με Χρήση του Reactome*

Πίνακας 4

Τυποποιημένα Επεξηγηματικά Σύνολα Δεδομένων για Εξαγωγή Σχέσεων

Σύνολο Δεδομένων	**Σύντομη Περιγραφή**
BioInfer[53]	Το BioInfer είναι ένα σώμα κειμένων (corpus) σε XML μορφή με αλληλεπιδράσεις πρωτεϊνών. Τα στοιχεία του BioInfer προήλθαν από πέντε γνωστά σώματα αλληλεπιδράσεων πρωτεϊνών: AIMED, BioInfer, LLL, IEPA και HPRD50.
HIV-1, human PI[54]	Το HIV-1 περιέχει περίληψη όλων των γνωστών αλληλεπιδράσεων πρωτεϊνών του HIV-1 με τις πρωτεΐνες του κύτταρου ξενιστή, διάφορες HIV-1 πρωτεΐνες, ή πρωτεΐνες από οργανισμούς που προέρχονται από ασθένειες που σχετίζονται με το HIV/AIDS.
LLL 05[55]	Το LLL05 αποτελείται από επεξηγηματικούς πράκτορες αναφοράς που στοχεύουν σε αλληλεπιδράσεις γονιδίων, από να λεξικό με ονοματικές οντότητες, καθώς και παραλλαγές και συνώνυμα και από γλωσσικές πληροφορίες. Το LLL05 μπορεί να χρησιμοποιηθεί για να αξιολογήσει την ικανότητα των συστημάτων για τον εντοπισμό των αλληλεπιδράσεων γονιδίων/πρωτεϊνών.
PICorpus[56]	Το PICorpus είναι ένα σώμα με αλληλεπιδράσεις πρωτεϊνών, το οποίο δημιουργήθηκε αρχικά στο PDG. Το PICorpus μπορεί να χρησιμοποιηθεί για μια ποικιλία από εργασίες εξόρυξης βιοατρικών

Σύνολο Δεδομένων	Σύντομη Περιγραφή
	κειμένων, όπως εξαγωγή ονοματικών οντοτήτων, αναγνώριση σχέσεων και εξαγωγή σχέσεων.
PDZBase[57]	Το PDZBase περιέχει 339 αλληλεπιδράσεις πρωτεϊνών του PDZ-τομέα, που έχουν εξαχθεί με μη αυτόματο τρόπο. Όλες οι αλληλεπιδράσεις έχουν προέλθει άμεσα από τον τομέα του PDZ, και είναι αναγνωρισμένες από in vivo ή in vitro πειράματα. Οι πληροφορίες από τις δεσμευτικές περιοχές αλληλεπίδρασης των πρωτεϊνών είναι γνωστές.
STRING[58]	Το STRING παρέχει γνωστές και προβλεπόμενες πρωτεϊνικές αλληλεπιδράσεις, συμπεριλαμβανομένων φυσικών και λειτουργικών ενώσεων που προέρχονται από τη γενετική, υψηλής απόδοσης πειράματα, συν-εκφράσεις και προηγούμενες γνώσεις.

Πίνακας 5

Συνήθη Χρησιμοποιούμενα Τυποποιημένα Επεξηγηματικά Σύνολα Δεδομένων για Εξόρυξη Κειμένων

Σύνολο Δεδομένων	Σύντομη Περιγραφή
BioCreative III[59]	Το BioCreative III λειτουργεί για αξιολόγηση της εξόρυξης κειμένων και συστημάτων εξαγωγής πληροφοριών που εφαρμόζονται στον τομέα της βιοιατρικής. Το BioCreative III έχει αρκετά σύνολα δεδομένων για τρεις εργασίες: Προσδιορισμό γονιδίων για διασταυρούμενα είδη και κανονικοποίηση, εξαγωγή πρωτεϊνικών αλληλεπιδράσεων και διαδραστικά εργαλεία επίδειξης για γονιδιακή ευρετηρίαση και ανάκτηση εργασιών.
BioInfer	Το BioInfer είναι ένα σώμα κειμένων (corpus) σε XML μορφή με αλληλεπιδράσεις πρωτεϊνών. Τα στοιχεία του BioInfer προήλθαν από πέντε γνωστά σώματα αλληλεπιδράσεων πρωτεϊνών: AIMED, BioInfer, LLL, IEPA και HPRD50.
BioText[60]	Το BioText αρχικά κατασκευάστηκε από 1000 τυχαίες επιλεγμένες περιλήψεις στο MEDLINE® από τα αποτελέσματα ενός ερωτήματος για τον όρο «μαγιά». Αυτό το σύνολο δεδομένων επεξηγήθηκε και επαληθεύτηκε περαιτέρω χειροκίνητα. Το BioText έχει 954 σωστά ζεύγη, συμπεριλαμβανομένων ορισμών συντομογραφιών, δεδομένα πρωτεϊνικών αλληλεπιδράσεων και σχέσεις μεταξύ οντοτήτων θεραπείας νόσων.
GENIA[61]	Το σύνολο δεδομένων GENIA είναι ένα από το πιο συχνά χρησιμοποιούμενα σύνολα δεδομένων για την αξιολόγηση των

Σύνολο Δεδομένων	Σύντομη Περιγραφή
	βιοιατρικών και βιολογικών πληροφοριών και συστημάτων εξόρυξης κειμένων. Το σύνολο δεδομένων περιέχει 1999 περιλήψεις από το Medline®, επιλεγμένες χρησιμοποιώντας ένα ερώτημα στο PubMed© για τους όρους «άνθρωπος», «κύτταρα αίματος» και «μεταγραφικοί παράγοντες». Το σύνολο δεδομένων GENIA έχει πολλά υποσύνολα δεδομένων, με στόχο την εν μέρει επεξήγηση, τη συντακτική επεξήγηση, επεξήγηση στην συντακτική ανάλυση (δομή μιας φράσης), την επεξήγηση όρων, γεγονότων, σχέσεων και συν-αναφορών.
PICorpus	Το PICorpus είναι ένα σώμα με αλληλεπιδράσεις πρωτεϊνών, το οποίο δημιουργήθηκε αρχικά στο PDG. Το PICorpus μπορεί να χρησιμοποιηθεί για μια ποικιλία από εργασίες εξόρυξης βιοατρικών κειμένων, όπως εξαγωγή ονοματικών οντοτήτων, αναγνώριση σχέσεων και εξαγωγή σχέσεων

Κεφάλαιο 3: Διατύπωση Επιστημονικών Υποθέσεων

3.1 Εισαγωγή

Οι βιολόγοι αντιμετωπίζουν τις εξακριβωμένες, ακόμα και τις διαψευσμένες επιστημονικές υποθέσεις ως ανώτερες των θεωρητικών μοντέλων, επειδή δίνουν περισσότερη αξία στα εμπειρικά στοιχεία. Ωστόσο, η αφθονία των ψηφιακών πληροφοριών, ειδικά στην μοριακή και κυτταρική βιολογία, είναι μια αναδυόμενη και πολλά υποσχόμενη πηγή πόρων για εννοιολογικές - θεωρητικές και μη εμπειρικές – βασισμένες σε βιβλιογραφία προσεγγίσεις για τη διατύπωση και δοκιμή υποθέσεων. Οι ψηφιακές βάσεις δεδομένων αντιπροσωπεύουν μια ευκαιρία για επιστημονική εξερεύνηση επειδή "ανακτήσιμα γεγονότα συγκεντρώνονται σε βάσεις δεδομένων από ποικίλες πηγές σε φαινομενικά άσχετα πεδία, καθώς και από τις χιλιάδες επιστημονικών περιοδικών"[62].

Η συνεχής ανάπτυξη εφαρμογών εξόρυξης γνώσης από κείμενα τις καθιστά χρήσιμες για την παραγωγή υποθέσεων και την εύρεση στοιχείων που να υποστηρίζουν τις υποθέσεις αυτές. Αρχικά, οι εφαρμογές αυτές διευκολύνουν εννοιολογικά πιο αποτελεσματικές ανακτήσεις – μια εξέλιξη που, για τους μελετητές που εκτίθενται σε ένα τεράστιο πλήθος πληροφοριών, είναι σίγουρα ευπρόσδεκτη. Τα εργαλεία αυτά μπορούν να γεφυρώσουν ασύνδετες βιβλιογραφίες άγνωστες σε εξειδικευμένους ερευνητές, ως απάντηση στην υπερφόρτωση πληροφοριών. Επιπλέον η σημερινή, τυπική τοπογραφία των δικτύων πληροφοριών χαρακτηρίζεται από κατευθυνόμενες συστάδες κόμβων, έτσι ώστε η αναζήτηση σε μια αποκαλούμενη «ήπειρο» να αποκλείει την αναζήτηση σε μια άλλη. Ως εκ τούτου, οι εφαρμογές εξόρυξης γνώσης μπορούν να γεφυρώσουν «ηπείρους» πληροφοριών στο διαδίκτυο και σε άλλα, χωρίς κλίμακα, δίκτυα. Είναι δυνατόν τέλος, να συγκρατήσουν την αλόγιστη σπατάλη πόρων μιας ψηφιακής βιβλιοθήκης, ενισχύοντας την πρόσβαση και προσθέτοντας αξία στο περιεχόμενο της.

Η δοκιμή των υποθέσεων στο πλαίσιο της εξόρυξης γνώσης από κείμενα αναφέρεται εν μέρει σε αυτοματοποιημένες διαδικασίες για την εύρεση στοιχείων που να υποστηρίζουν υποθετικές σχέσεις. Ένας σημαντικός στόχος των πληροφορικών σε συνεργασία με ειδικούς διαφόρων τομέων, είναι η εξαγωγή αρκετών στοιχείων που θα υποστηρίζουν υποθέσεις που ενδιαφέρουν τους εμπειριστές για ενδεχόμενη πειραματική επαλήθευση.

3.2 Ορισμός

Οι ερευνητικές υποθέσεις είναι το επίκεντρο των επιστημονικών αναζητήσεων. Η ακριβής, σαφής και λειτουργική αναπαράσταση τους είναι ζωτικής σημασίας για την επίσημη καταγραφή και ανάλυση των επιστημονικών ερευνών. Οι υποθέσεις πρέπει να αναπαριστώνται και να καταγράφονται έτσι ώστε να συλλάβουν τη σημειολογική σημασία της υπόθεσης και να προωθήσουν τον χειροκίνητο (ή αυτοματοποιημένο) σχεδιασμό πειραμάτων για να δοκιμαστούν αυτές οι υποθέσεις.

Ενώ η εξόρυξη σχέσεων επικεντρώνεται στην εξαγωγή των σχέσεων μεταξύ των οντοτήτων που περιλαμβάνονται ρητά στο κείμενο, η ***διατύπωση υποθέσεων*** προσπαθεί να αποκαλύψει τις σχέσεις που δεν υπάρχουν στο κείμενο, αλλά αντιθέτως έχουν συναχθεί από

την παρουσία άλλων, πιο ρητών σχέσεων. Ο στόχος είναι να αποκαλύψει προηγουμένως παραγνωρισμένες σχέσεις αντάξιες για περαιτέρω έρευνα.

Μια ***υπόθεση*** είναι μια εικαστική δήλωση σχετικά με τη σχέση μεταξύ δύο ή περισσοτέρων μεταβλητών[63]. Υπάρχουν δύο βασικά χαρακτηριστικά που πρέπει να έχουν όλες οι υποθέσεις: πρέπει να είναι δηλώσεις της σχέσης μεταξύ των μεταβλητών και πρέπει να φέρουν σαφείς ενδείξεις για τον έλεγχο των δηλωμένων σχέσεων. Αυτά τα χαρακτηριστικά συνεπάγονται ότι πρόκειται για τις σχέσεις και όχι για τις μεταβλητές, οι οποίες τίθενται υπό έλεγχο. Οι υποθέσεις καθορίζουν πώς συνδέονται οι μεταβλητές και ότι αυτές είναι υπολογίσιμες ή δυνητικά υπολογίσιμες. Οι προτάσεις που δεν έχουν κάποιο ή όλα αυτά τα χαρακτηριστικά δεν είναι υποθέσεις έρευνας. Για παράδειγμα, έστω ότι γίνεται η εξής υπόθεση:

«Τα επίπεδα σιδήρου στον οργανισμό αυξάνονται καθώς αυξάνεται η κατανάλωση κόκκινου κρέατος»

Πρόκειται για μια σχέση που δηλώνεται ανάμεσα σε μία μεταβλητή, "*κατανάλωση κόκκινου κρέατος*" και μια άλλη μεταβλητή, "*επίπεδα σιδήρου*". Επιπλέον, οι δύο μεταβλητές είναι δυνητικά μετρήσιμες, οπότε τα κριτήρια έχουν εκπληρωθεί. Ωστόσο για τους σκοπούς των στατιστικών δοκιμών είναι πιο σύνηθες να βρεθούν υποθέσεις που δηλώνονται με τη λεγόμενη κενή μορφή, όπως φαίνεται παρακάτω:

«Δεν υπάρχει καμία σχέση μεταξύ κατανάλωσης κόκκινου κρέατος και των επιπέδων σιδήρου στον οργανισμό».

Οι υποθέσεις έχουν κεντρική σημασία για την πρόοδο στην έρευνα. Κατευθύνουν τις προσπάθειες των ερευνητών, επιβάλλοντάς τους να επικεντρωθούν στη συγκέντρωση γεγονότων ή δεδομένων, που θα τους επιτρέψει να εξακριβώσουν ή όχι τις υποθέσεις. Υπάρχει ένα δεύτερο πλεονέκτημα στην διατύπωση υποθέσεων, ότι σιωπηρές έννοιες ή εξηγήσεις για διάφορα γεγονότα γίνονται ρητές και αυτό συχνά οδηγεί σε τροποποιήσεις των δηλώσεων αυτών, ακόμη και πριν συλλεχθούν τα δεδομένα.

Περιστασιακά, μια δεδομένη υπόθεση μπορεί να είναι αρκετά ευρεία ώστε να δοκιμαστεί. Ωστόσο, άλλες υποθέσεις υπό έλεγχο μπορούν να προκύψουν από αυτή. Ένα πρόβλημα δεν μπορεί να λυθεί κυριολεκτικά, εάν δεν είναι μειωμένο σε μορφή υποθέσεως, επειδή ένα πρόβλημα είναι ένα ερώτημα, συνήθως ευρείας φύσεως και δεν είναι άμεσα ελέγξιμο.

3.3 Διατύπωση Υποθέσεων στη Βιοιατρική

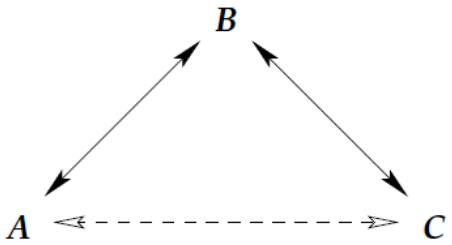
Ο Swanson[64] ήταν ένας από τους πρώτους υποστηρικτές της διατύπωσης υποθέσεων από τη βιοιατρική βιβλιογραφία. Η «άγνωστη δημόσια γνώση», φράση που επινοήθηκε από τον Swanson, αναφέρεται στη δημοσιευμένη γνώση που είναι αποτελεσματικά «θαμμένη» σε ασύνδετους θεματικούς τομείς – «ασύνδετους», διότι οι ερευνητές που εργάζονται σε διάφορα πεδία, αγνοούν ο ένας τον άλλον. Ως εκ τούτου, πραγματικά ασύνδετες βιβλιογραφίες δεν έχουν κανένα άρθρο από κοινού. Ο Swanson πρότεινε σε μια σειρά από καινοτόμα έγγραφα ότι καινούριες πληροφορίες μπορεί να ανακαλυφθούν από συστηματική μελέτη φαινομενικά άσχετων και μη αλληλεπιδραστικών ερευνητικών βιβλιογραφιών, τις οποίες ονόμασε «συμπληρωματικές αλλά ασύνδετες». Για να αποδείξει την σκοπιμότητα των ιδεών του και ως αποτέλεσμα του σημαντικού του έργου, ο Swanson

ανακάλυψε μια νέα σύνδεση μεταξύ της νόσου Raynaud και του ιχθυέλαιου, εξετάζοντας δύο αταίριαστα σύνολα βιοιατρικής βιβλιογραφίας. Η υπόθεση των ευεργετικών επιδράσεων του ιχθυέλαιου σχετικά με τη νόσο του Raynaud επιβεβαιώθηκε από ανεξάρτητες κλινικές δοκιμές δύο χρόνια αργότερα, γεγονός που απέδειξε την αξία της εξόρυξης γνώσης από βιοιατρική λογοτεχνία σε επιστημονικές ανακαλύψεις.

3.3.1 Το μοντέλο υποθέσεων ABC

Το υποθετικό μοντέλο του Swanson, αποκαλούμενο και ως ***Swanson's ABC model***, μπορεί να περιγραφθεί απλά ως εξής:

" το A σχετίζεται με το B, το B σχετίζεται με το C, άρα το A μπορεί να σχετίζεται με το C " [65]



Εικόνα 8 Το μοντέλο ανακάλυψης ABC του Swanson.

Οι σχέσεις AB και BC είναι γνωστές και αναφέρονται στη βιβλιογραφία. Η έμμεση σχέση AC είναι πιθανόν μια νέα ανακάλυψη.

Μια συνοπτική περίληψη του μοντέλου αυτού είναι η παρακάτω:

Δοθείσης μια συγκεκριμένης ερώτησης έρευνας στη βιοιατρική, ένας αρχικός στόχος είναι να εντοπιστούν δύο συμπληρωματικές, αλλά ασύνδετες, βιβλιογραφίες, έστω **AB** και **BC**, όπου **A**, **B** και **C** είναι μεταβλητές ή έννοιες που έχουν ενδιαφέρον. Αρχίζοντας με την αναζήτηση τίτλων στο **MEDLINE®** σχετικές με το **C** και στη συνέχεια με το **A**, εξετάζονται τα αποτελέσματα και δημιουργείται μια λίστα των τίτλων με τον κοινόχρηστο όρο **B**. Λαμβάνοντας από κοινού τους δυο όρους, δηλαδή τους **AB** και **BC**, αυτοί είναι ανεξάρτητοι, δεδομένου ότι τίποτα δεν έχει δημοσιευθεί που να συνδέει το **A** με το **C**. Για παράδειγμα, έστω ότι το **C** αντιπροσωπεύει πηγές βιβλιογραφίας σχετικές με την ημικρανία και το **A** την στοχευμένη βιβλιογραφία που αφορά το μαγνήσιο. Το **B** είναι η ενδιάμεση λογοτεχνία που χρησιμοποιείται ως σύνδεση του **A** με το **C**. Μετά από επανεξέταση εμπειρογνωμόνων, η κοινή λίστα των όρων **B** στους τίτλους της **AB** και **BC**, τελικά οδηγεί σε αρκετές νέες υποθέσεις για έλεγχο, σχετικά με τις φυσιολογικές επιπτώσεις της ανεπάρκειας μαγνησίου σε σχέση με την ημικρανία. Σε αυτό το σημείο, έστω και αν ένα σύνολο υποτιθέμενων σχέσεων έχει ανακαλυφθεί, απαιτούνται ανεξάρτητες πειραματικές δοκιμές για την επικύρωση των αποτελεσμάτων, όπως για παράδειγμα διεξαγωγή κλινικών δοκιμών.

Για την περαιτέρω αυτοματοποίηση της μεθόδου αυτής, οι Swanson και Smallheiser, ανέπτυξαν ένα διαδραστικό λογισμικό που ονομάζεται **ARROWSMITH**[66] και είναι διαθέσιμο στο διαδίκτυο. Στο site, ο χρήστης επιλέγει μία από δύο λειτουργίες (διατύπωση υπόθεσης ή έλεγχο υπόθεσης) ώστε να παράγει μια λίστα **A** και μια λίστα **C** των όρων, κάνοντας αναζήτηση σε τίτλους στο **MEDLINE®** και ιατρικά υπαγόμενους τίτλους μέσω του **PubMed©** ή του **OVID**[67]. Η λειτουργία ελέγχου υποθέσεων αίρει την αρχική υπόθεση για καθαρά ασύμβατα ζεύγη βιβλιογραφίας, δεδομένου ότι από τη στιγμή που υπάρχει μια πιθανή σχέση, άρθρα που παραπέμπουν στα **A**, **B**, και **C** πιθανώς να υπάρχουν αλλά δεν είναι κοινώς γνωστά.

Επιπλέον, οι Swanson και Smallheiser αναγνώρισαν το γεγονός ότι δύο λογοτεχνίες μπορεί να συνδέονται ψευδώς, λόγω κοινής γλώσσας σαν ευρύτερη αρχή, λόγου χάρη της ιατρικής. Για τον λόγο αυτό, όρισαν φίλτρα στην πρώιμη έκδοση του **ARROWSMITH** που έλεγχαν αυτήν την πιθανή σύγχυση και εισήγαγαν ανθρώπινη νοημοσύνη στο διαδραστικό σύστημα. Στα πρώτα φίλτρα, συμπεριλαμβανόταν μια λίστα με a priori διακόπτουσες λέξεις (υλοποιημένη χειροκίνητα και όχι με τη βοήθεια υπολογιστή), ένα στατιστικό "κατώφλι" για τη διατήρηση όρων με βάση τη σχετική συχνότητα, καθώς και περιορισμοί κατηγοριών, για παράδειγμα «διατροφικοί παράγοντες» ή «τοξίνη». Η πρόσφατη έκδοση του **ARROWSMITH** προσφέρει πρόσθετα φίλτρα, όπως «πρώτη ημερομηνία δημοσίευσης».

3.3.2 Μελέτες Swanson Linking και Ανάπτυξη

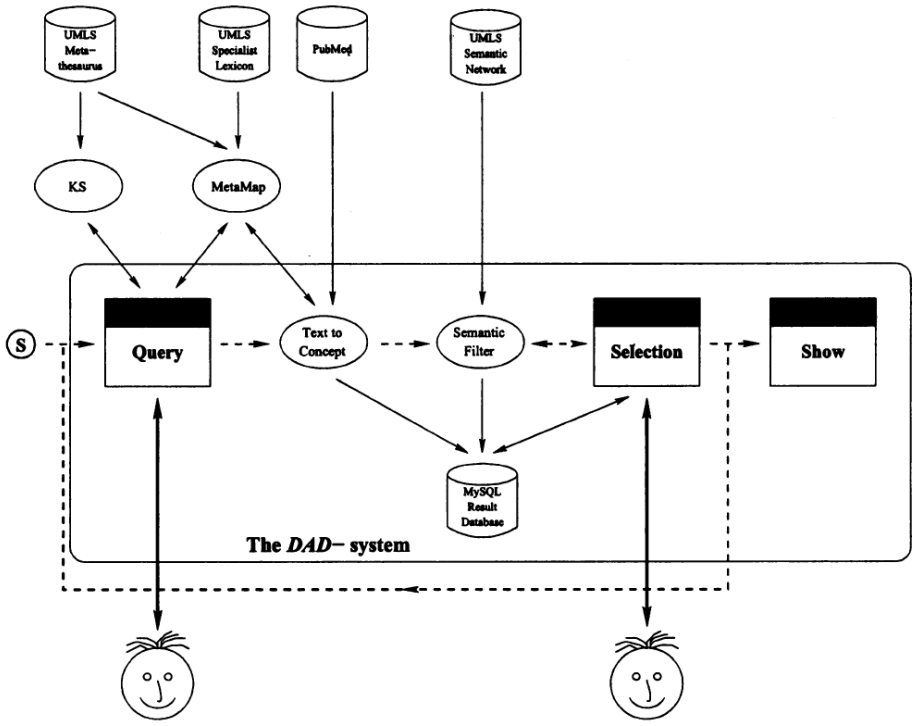
Ο όρος ***Swanson Linking (SL)***, αναφέρεται σε «*ανακαλύψεις βασισμένες σε βιβλιογραφία, στις οποίες ο **SL** μπορεί να οριστεί ως η εξεύρεση ασύνδετων βιβλιογραφικών συνδέσμων, καθιερώνοντας νοηματικές συνδέσεις μεταξύ τους, με χρήση ανάκτησης πληροφοριών από βιβλιογραφικές βάσεις δεδομένων*». Οι ερευνητές που συνέχισαν το έργο του Swanson, παρέμεναν πιστοί στη λογική αυτή, αλλά ίσως έδειξαν παραπάνω σεβασμό στις μεθόδους που χρησιμοποίησε. Για παράδειγμα, οι υποθέσεις στις οποίες βασίστηκαν οι μελέτες SL, περιλαμβάνουν συνήθως μια ασθένεια, η βάση επιλογής είναι η **MEDLINE®** και η αξιολόγηση σχεδόν πάντα περιλαμβάνει αναπαραγωγή των αρχικών ευρημάτων του Swanson. Ακόμα και έτσι όμως, οι ερευνητές έχουν κάνει ανεκτίμητες συνεισφορές με τη συστηματοποίηση των πρόωρων μεθόδων του Swanson, τη βελτίωση της αυτοματοποίησης ορισμένων πτυχών της διατύπωσης υποθέσεων και την εξόρυξη οντοτήτων πέρα από τους τίτλους.

Παρακάτω γίνεται αναφορά στις σημαντικότερες μελέτες SL που βοήθησαν στην ανάπτυξη του τομέα της Διατύπωσης Υποθέσεων.

3.3.2.1 Gordon and Lindsay

Το 1996, οι Gordon και Lindsay[68] δημοσίευσαν μια μελέτη σχετικά με τα συστήματα υποστήριξης ανακάλυψης γνώσης, επειδή «*κανένας άλλος ερευνητής δεν είχε αναφέρει τη διεξαγωγή πειραμάτων που ανακαλύφθηκαν βασισμένα στη βιβλιογραφία, που να επιβεβαιώνουν, ή να αναιρούν σε έκταση το έργο του Swanson με κάθε τρόπο*». Αυτό ήταν μια δεκαετία μετά την πρώτη δημοσίευση του Swanson πάνω στην εξόρυξη γνώσης από κείμενα. Τα αποτελέσματά τους έδωσαν πιστότητα στην στρατηγική που ακολούθησε ο

Swanson, επιβεβαιώνοντας τη σύνδεση μεταξύ του συνδρόμου Raynaud και του ιχθυέλαιου. Επιπλέον όμως, εισήγαγαν λεξιλογικές και στατιστικές μεθόδους για την εξόρυξη περιλήψεων αντί τίτλων και ανέπτυξαν υπολογιστικά εργαλεία για την υποστήριξη ανακάλυψης γνώσης. Συγκρίνοντας διάφορα μέτρα συχνότητας για την επιλογή όρων, εισήγαγαν την ***ποσοτική αυστηρότητα*** στο πεδίο της διατύπωσης υποθέσεων.



Εικόνα 9 Το σύστημα DAD.

Το οβάλ πλαίσιο αναπαριστά το σύστημα αυτό καθαυτό, ενώ οτιδήποτε εκτός πλαισίου αποτελεί πηγή.

3.3.2.2 Weeber et al.

Ο Weeber[69] και οι συνεργάτες του ανέπτυξαν ένα εννοιολογικά βασισμένο σύστημα επεξεργασίας φυσικής γλώσσας (natural language processing system) που ονομάζεται **DAD** (**D**rug - **A**dverse Drug Reaction – **D**isease) (**Εικόνα 9**) με σκοπό να βοηθήσει εμπειρογνώμονες της βιοιατρικής στη διατύπωση και τον έλεγχο υποθέσεων, κυρίως για έρευνες που αφορούν ανακάλυψη φαρμάκων. Η δυσκολία εξαγωγής λέξεων παρακάμφθηκε, καταργώντας την ανάγκη για λίστες διακοπτουσών λέξεων και σύνθετων ερωτημάτων για συνώνυμα και παραλλαγές και αντιστοιχίζοντας λέξεις σε τίτλους και περιλήψεις με έννοιες από το **UMLS®** **Metathesaurus®**[70], μια από τις τρεις συνιστώσες της Εθνικής Βιβλιοθήκης Ιατρικής **UMLS®**[71]. Η χαρτογράφηση αυτού του είδους, διευκολύνει την εξόρυξη σύνθετων φράσεων, όπως «πίεση αίματος» και περιορίζει τον χώρο αναζήτησης, χρησιμοποιώντας σημασιολογικούς τύπους της **UMLS** ως φίλτρα. Μέχρι και τον Μάρτιο του 2013, το σημασιολογικό δίκτυο του **UMLS** περιέχει 133 σημασιολογικές κατηγορίες και 54 σημασιολογικές σχέσεις, που

προσδιορίζουν τη δομή του δικτύου και αντιπροσωπεύουν σημαντικές σχέσεις στον βιοιατρικό τομέα[72].

Για να αποδείξει τη χρησιμότητα του συστήματος υποθέσεων τους, ο Weeber και οι συνεργάτες του δημοσίευσαν τα αποτελέσματα μιας ενδιαφέρουσας μελέτης για ενδεχομένως νέους στόχους ασθενειών για το φάρμακο θαλιδομίδη[73]. Βρήκαν βιβλιογραφικά στοιχεία στο **PubMed©**, γεγονός που υποδηλώνει ότι η θαλιδομίδη θα μπορούσε να είναι μια αποτελεσματική θεραπεία για ασθένειες όπως η χρόνια ηπατίτιδα C, βαριά μυασθένεια, γαστρίτιδα που προκαλείται από το βακτήριο *Helicobacter-pylori* και η οξεία παγκρεατίτιδα.

3.3.2.3 Stegmann and Grohmann

Οι Stegmann και Grohmann[74] επέκτειναν την μεθοδολογία SL, με την χρησιμοποίηση ανάλυσης συν-λέξεων (***co-word analysis***), μια στατιστική μέθοδο, που χρησιμοποιείται στη συσταδοποίηση. Αντί για λέξεις ή έννοιες, ανέλυσαν τη δύναμη συν-εμφάνισης στα σύνολα ανάκτησης, διαφόρων ζευγών από λέξεις-κλειδιά, που ανέθεσαν σε έγγραφα του **MEDLINE®**. Στις λέξεις-κλειδιά περιλαμβάνονται τίτλοι ιατρικών θεμάτων (**MeSH**), το **EC_number** (*Enzyme Commission number*, μια μέθοδος ταξινόμησης των ενζύμων) καθώς και αριθμοί μητρώων **CAS** (Chemical Abstracts Service, παράρτημα της Αμερικανικής Ένωσης Χημικών **ACS**). Οι αναλύσεις οδηγούν σε χάρτες ή "στρατηγικά διαγράμματα" συμπλεγμάτων που περιέχουν τις λέξεις-κλειδιά. Πολλά υποσχόμενες συνδέσεις όρων **B** (από το μοντέλο ABC του Swanson) που οδηγούν σε συσχέτιση τις ασύνδετες βιβλιογραφίες **A** και **C**, τείνουν να εμφανίζονται σε περιοχές χαμηλής κεντρικότητας και πυκνότητας. Η προσέγγιση τους αυτή, επικυρώθηκε από την αναπαραγωγή των ευρημάτων του Swanson για το σύνδρομο Raynaud σε σχέση με το ιχθυέλαιο και για την ημικρανία σε σχέση με το μαγνήσιο. Βρήκαν επίσης αποδεικτικά στοιχεία για σχέση μεταξύ prions (το prion είναι ένας μολυσματικός παράγοντας που αποτελείται κυρίως από πρωτεΐνες), νευροεκφυλιστικών ασθενειών και μαγγανίου. Αυτή η σχέση είχε χαρτογραφηθεί νωρίτερα από τον Chen στο πλαίσιο λανθανόντων τομέων γνώσης – «λανθάνων» λόγω του χαμηλού ποσοστού παραπομπής που έχει ένα σημαντικό έγγραφο για χαρτογράφηση, όπως περιγράφεται στο βιβλίο του Chen[75].

Ένα πλεονέκτημα της ανάλυσης συν-λέξεων και της συσταδοποίησης είναι ότι τα πρώιμα στάδια της επιλογής όρων είναι αυτοματοποιημένα. Ωστόσο, οι εμπειρογνώμονες από κάποιον τομέα, πρέπει ακόμα και έτσι να επανεξετάσουν τις συστάδες για την τελική επιλογή των κατάλληλων όρων. Ένα άλλο πλεονέκτημα είναι ότι μπορεί να είναι ευκολότερο για τους χρήστες να επανεξετάσουν χάρτες ή διαγράμματα συστάδων παρά μεγάλες λίστες ταξινομημένων όρων. Ένα μειονέκτημα που παρουσιάζει η μέθοδος είναι ότι εξαρτάται από λέξεις-κλειδιά που προέρχονται από ένα ελεγχόμενο λεξιλόγιο. Άλλες μέθοδοι όπως η εξόρυξη τίτλων και περιλήψεων, είναι πιο κατάλληλες εάν λείπουν κάποιες λέξεις-κλειδιά.

3.3.2.4 Srinivasan

Η προσέγγιση που ακολούθησε η Srinivasan για τον εντοπισμό των πολλά υποσχόμενων όρων **B**, ξεκινά με την οικοδόμηση δύο κατατομών για τα θέματα **A** και **C** αντίστοιχα, από τα σύνολα αποτελεσμάτων που ανακτήθηκαν για το **A** και το **C**[76]. Στο έργο της, η κατατομή ενός

θέματος αποτελείται από όρους, που έχουν υψηλή συχνότητα εμφάνισης στο σύνολο ανακτηθέντων αποτελεσμάτων του εκάστοτε θέματος και ανήκουν σε σημασιολογικές κατηγορίες που ενδιαφέρουν τον χρήστη. Τότε η τομή της κατατομής του **A** με αυτήν του **C**, δημιουργεί τους υποψήφιους **B** όρους. Η διαδικασία αναγνώρισης **B** όρων από δεδομένα θέματα **A** και **C** αποκαλείται ***κλειστή ανακάλυψη***. Η Srinivasan εισήγαγε επίσης την ιδέα της θεματικής κατατομής για διεξαγωγή ανοιχτής ανακάλυψης, η οποία αναγνωρίζει **B** και **C** όρους, έχοντας ως μόνο δεδομένο το θέμα **A**.

Ο αναφερόμενος αλγόριθμος ανοιχτής ανακάλυψης μπορεί να περιγραφεί απλά ως εξής: κορυφαίοι όροι **B** επιλέγονται από την κατατομή του θέματος **A**. Στη συνέχεια, δημιουργείται μια κατατομή για κάθε επιλεγμένο όρο **B**, από το σύνολο των ανακτηθέντων αποτελεσμάτων για τον συγκεκριμένο όρο **B**. Οι κορυφαίοι όροι στην κατατομή του θέματος **B** καθιστούν υποψήφιους όρους **C**. Εάν το σύνολο των ανακτηθέντων αποτελεσμάτων του θέματος **A** είναι ασύνδετο με το σύνολο ανακτηθέντων αποτελεσμάτων ενός υποψηφίου όρου **C**, τότε αυτός ο όρος **C** αναφέρεται ως έχων πιθανή σχέση με το θέμα **A** μέσω του όρου **B**.

Η Srinivasan δημοσίευσε τα αποτελέσματα μια εκτεταμένης αναπαραγωγής των μελετών του Swanson και Smalheiser, προσεκτικά συγκρίνοντας τις μεθόδους της με τη δική τους, καθώς και με αυτές των Gordon και Lindsay και του Weeber, με τη χρήση ενός ενεργού ερευνητικού προγράμματος εξόρυξης γνώσης από κείμενα. Σε μια από τις εφαρμογές του προγράμματος της, όπου επιδεικνύεται η χρησιμότητα του, διερευνώνται τα θεραπευτικά οφέλη της *Curcuma longa* (κουρκουμίνη) για ασθένειες του αμφιβληστροειδούς, την νόσο του Crohn και διαταραχές του νωτιαίου μυελού[77]. Το έργο της μοιάζει με αυτά των Weeber και των Stegmann – Grohmann, στο ότι χρησιμοποιεί σημασιολογικούς τύπους από το **UMLS** και μετα-δεδομένα από το **MEDLINE®** (όροι **MeSH**) αντίστοιχα, ωστόσο συνδυάζει αυτά τα στοιχεία με ένα τρόπο πολύ διαφορετικό από την κάθε ομάδα.

Οι αλγόριθμοι εξόρυξης γνώσης από κείμενα της Srinivasan περιλαμβάνουν δομικές κατατομές ερευνητικών θεμάτων, με βάση σταθμισμένους όρους **MeSH**, οι οποίοι ανακτήθηκαν από έγγραφα του **MEDLINE®**, τα βάρη των οποίων εκτιμώνται βάσει σημασιολογικών τύπων. Αν ληφθούν υπόψη όλα αυτά μαζί, οι σταθμισμένοι όροι συγκροτούν μια κατατομή του θέματος προς μελέτη. Για παράδειγμα, μα κατατομή για το κληρονομικό σύνδρομο Marfans, κατά πάσα πιθανότητα θα αποτελείται σε μεγάλο βαθμό από σταθμισμένους όρους όπως «γονίδια, πρωτεΐνες, συμπτώματα, φαρμακευτικές αγωγές, άλλες ασθένειες και ομάδες πληθυσμού». Τα θέματα για κατατομή, μπορεί να είναι μεμονωμένες λέξεις ή φράσεις που δεν χρειάζεται να είναι ή να αποτελούνται από όρους **MeSH**. Σε αντίθεση με τα αποτελέσματα των Stegmann-Grohmann, τα αποτελέσματα είναι κατανεμημένες λίστες όρων και όχι συστάδες.

Ελαφρώς διαφορετική από την προσέγγιση που ακολούθησε η Srinivasan, ήταν η διεξαγωγή ανοιχτής ανακάλυψης που ακολούθησαν οι Pratt-Yildiz[78], εφαρμόζοντας απευθείας εξόρυξη σχέσεων στο σύνολο αποτελεσμάτων του θέματος **A**. Στην μελέτη τους, η λογική συμπερασματολογία, που βασίζεται σε δύο κανόνες συσχέτισης **A** → **B** και **B** → **C**, οδηγεί στην εύρεση ενός υποψήφιου όρου **C**.

3.4 Πρόσφατες εξελίξεις

Πέρα από την ανακάλυψη του Swanson, αρκετή έρευνα έχει πραγματοποιηθεί με στόχο την αυτοματοποίηση και τον εξευγενισμό του μοντέλου ανακάλυψης ABC του Swanson. Ωστόσο, οι περισσότερες από τις αναφερόμενες προσεγγίσεις βασίζονται στην ανάλυση του συνόλου ανακτηθέντων αποτελεσμάτων για ένα ή δύο αρχικά θέματα, που παρέχονται ως ερώτημα από ένα χρήστη, αντί να είναι σε θέση να αναβαθμιστούν στο σύνολο της βάσης δεδομένων της βιβλιογραφίας για το σκοπό της ανακάλυψης καινοτόμων βιοιατρικών υποθέσεων, διασταυρωμένων με πολλαπλές αποθήκες δεδομένων.

Αν μοντελοποιηθεί μια αποθήκη βιοιατρικής βιβλιογραφίας ως ένα ολοκληρωμένο δίκτυο βιοιατρικών εννοιών που ανήκουν σε διαφορετικούς σημασιολογικούς τύπους, οι τεχνικές ανακάλυψης συνδέσεων μπορούν να επιτρέψουν την ανακάλυψη υποθέσεων μεγάλης κλίμακας, διασταυρωμένων με πολλές αποθήκες δεδομένων, ξεπερνώντας τις ανακαλύψεις που βασίζονται στην ανάκτηση πληροφοριών[79]. Τα τελευταία χρόνια, έχει μελετηθεί εκτενώς η ανακάλυψη συνδέσεων – σχέσεων από κοινωνικά δίκτυα, όπως οι σχέσεις που λαμβάνονται από δεδομένα από το Facebook® και βιβλιογραφικές βάσεις δεδομένων που ανήκουν στο DBLP[80]. Ως ένα σημαντικό πρόβλημα της εξόρυξης συνδέσεων, η ανακάλυψη συνδέσεων αναφέρεται στην ανακάλυψη των μελλοντικών σχέσεων μεταξύ αντικειμένων (ή κόμβων) που δεν συνδέονται άμεσα με το τρέχον στιγμιότυπο ενός συγκεκριμένου δικτύου. Σε πρόσφατη έρευνα[81], εφαρμόστηκε μια τεχνική ανακάλυψης συνδέσεων, για τη διατύπωση υποθέσεων για τις σχέσεις μεταξύ των γονιδίων και των εμβολίων. Στην έρευνα αυτή, εξήχθησαν πρώτα δίκτυα για τις γονιδιακές αλληλεπιδράσεις και τις αλληλεπιδράσεις γονιδίων-εμβολίων από λογοτεχνία με τη βοήθεια οντολογιών γονιδίων και εμβολίων και στη συνέχεια αναλύθηκαν τα δίκτυα με τον υπολογισμό διαφόρων τύπων για μέτρα κεντρικότητας του κάθε κόμβου σε αυτά τα δίκτυα. Δεδομένης της περιορισμένης εστίασης όμως στις σχέσεις γονιδίων και εμβολίων, η έρευνα αυτή από τη φύση της δεν είχε σχεδιαστεί για βιοιατρικές ανακαλύψεις που να διασταυρώνονται με αποθήκες δεδομένων.

Μια καινούρια προσέγγιση είναι αυτή του συνδυασμού της εξόρυξης γνώσης από κείμενα και των οντολογιών για την διατύπωση υποθέσεων[82]. Σε αυτήν την προσέγγιση, τα εργαλεία εξόρυξης γνώσης από κείμενα εφαρμόζουν αναγνώριση οντοτήτων ώστε να ανακαλύψουν όρους που έχουν επιλεγεί από μια οντολογία. Η γνώση που κρύβεται πίσω από τις οντολογίες — για παράδειγμα, οι σχέσεις μεταξύ των όρων και ειδικότερα η ιεραρχία των όρων — μπορεί να είναι χρησιμοποιηθεί για την κατηγοριοποίηση των χωρίων του κειμένου σύμφωνα με μια οντολογική ιεραρχία ή να προσδιοριστούν δηλώσεις που αντιπροσωπεύουν περιπτώσεις οντολογικής γνώσης. Για την διατύπωση υποθέσεων, τα αποτελέσματα από διαφορετικούς πόρους (για παράδειγμα, βιβλιογραφία και βάσεις δεδομένων) μπορεί να συγκριθούν το ένα έναντι του άλλου χρησιμοποιώντας κριτήρια σημασιολογικής ομοιότητας[83]. Οι διαφορές μεταξύ των πόρων ή των μορφών των αποδεικτικών στοιχείων μπορούν να χρησιμοποιηθούν για να βελτιωθεί η αναπαράσταση της εκάστοτε οντολογικής γνώσης.

Για παράδειγμα, η εξόρυξη γνώσης από κείμενα έχει χρησιμοποιηθεί για να προσδιορίσει παρατηρήσεις από αλληλεπιδράσεις μεταξύ γονιδίων, φαρμάκων και φαινότυπων στη βιβλιογραφία. Οι εξαγόμενες σχέσεις (όπως μεταβολισμός, αναστολή και ενεργοποίηση) χαρτογραφήθηκαν σε μια κοινή οντολογία και στη συνέχεια αλληλεπιδράσεις μεταξύ φαρμάκων μπορούσαν να προσδιοριστούν με την χρήση μια μηχανής μάθησης[84,85]. Κάθε

προβλεπόμενη αλληλεπίδραση μεταξύ φαρμάκων, συνδέεται με συγκεκριμένες δηλώσεις που υποστηρίζουν και εξηγούν την υποθετική αλληλεπίδραση. Σε μια άλλη μελέτη[86], προσδιορίστηκαν όροι από μια οντολογία φαινότυπου για ετικέτες φαρμάκων και χρησιμοποιήθηκαν τα αποτελέσματα για τον ορισμό ενός μέτρου ομοιότητας μεταξύ των παρενεργειών. Το έργο αυτό έδειξε ότι η ομοιότητα ανάμεσα σε παρενέργειες φαρμάκων, μπορεί να χρησιμοποιηθεί για την υπόθεση νέων στόχων για τα φάρμακα και νέες αλληλεπιδράσεις μεταξύ των φαρμάκων – 13 αναμενόμενοι στόχοι φαρμάκων και 20 νέες αλληλεπιδράσεις φαρμάκων επικυρώθηκαν πειραματικά από in vitro δεσμευτικές δοκιμές.

3.5 Εργαλεία για Διατύπωση Υποθέσεων

Υπάρχει μια σειρά από προγράμματα που αποσκοπούν στην αντιμετώπιση της ανάγκης να αναπαρασταθούν και να καταγραφούν οι ερευνητικές υποθέσεις σε σημασιολογικά καθορισμένη μορφή. Οι υποθέσεις στη γνωσιακή βάση για τη νόσο του Alzheimer **SWAN**[87] (Semantic Web Applications in Neuroscience), αποτελούν τμήματα κειμένων φυσικής γλώσσας, που αναπαρίστανται ως δηλώσεις έρευνας (στοιχεία λόγου), τα οποία συνδέονται (μέσω λόγου-σχέσεων) με άλλα στοιχεία λόγου και αναφορές που καθορίζουν το όνομα του συγγραφέα, το άρθρο, το περιοδικό κ.α.[88]. Ομοίως, το **OBI**[89] (Ontology for Biomedical Investigations) προτυποποιεί τις υποθέσεις ως την τάξη *obi:hypothesis textual entity*, όπου οι υποθέσεις είναι μέρος της *obi: objective specification* του *obi:investigation*. Το πρόγραμμα **ART**[90] θεωρεί τις επιστημονικές εργασίες ως αναπαράσταση κειμένου σε επιστημονικές έρευνες και χρησιμοποιεί τις τάξεις-κλειδιά από την γενική οντολογία των πειραμάτων **EXPO**[91] για να προσθέσει σχόλια στα έγγραφα. Η κλάση *expo:hypothesis* χρησιμοποιείται για να επισημάνει προτάσεις που περιγράφουν υποθέσεις έρευνας. Για παράδειγμα, το έγγραφο *b310850* από το Corpus του **ART** περιέχει μια πρόταση που έχει ήδη επισημανθεί ως μια υπόθεση:

<s sid="41"><annotationART atype="GSC" type="Hyp" conceptID="Hyp1" novelty="-None" advantage="None">This means that whereas a central ligand may change chemical properties somewhat, this should only be a second order effect on the properties we are studying here.</annotationART></s>

Η εξόρυξη υποθέσεων από τη βιβλιογραφία ως οντότητες κειμένου, καθώς και η απόθεση αυτών των υποθέσεων σε κοινά διαθέσιμες, περιεκτικές και σημασιολογικά επισημασμένες συλλογές, ανοίγει νέες προοπτικές για την ανταλλαγή και διαμοίραση γνώσεων. Η ανοιχτή και εύκολη πρόσβαση σε μια σειρά από εναλλακτικές υποθέσεις που αντικατοπτρίζουν μια πληθώρα συχνά συγκρουόμενων θεωριών, αντιλήψεων και απόψεων θα μπορούσε να επιταχύνει σημαντικά την πρόοδο της επιστήμης. Παρόλα αυτά, είναι συνήθως δύσκολο να συλλάβει κανείς την ακριβή σημειολογική σημασία μιας υπόθεσης που εκφράζεται ως οντότητα κειμένου. Μερικές φορές είναι αδύνατο να κατανοηθεί η έννοια και να επεξεργαστεί σωστά μια υπόθεση χωρίς την ανάγνωση μιας σημαντικής μερίδας του περιβαλλόμενου κειμένου. Η έγγραφη αναπαράσταση των υποθέσεων που ανακτήθηκαν από τη βιβλιογραφία προορίζεται ως επί το πλείστον για ανθρώπινη κατανόηση και έχει περιορισμένη αξία για την αυτόματη επεξεργασία.

Ένας αριθμός προγραμμάτων προσπάθησε να ξεπεράσει αυτόν τον περιορισμό και να μεταφράσει υποθέσεις σε μια επεξεργάσιμη από μηχανές μορφή. Το **HyBrow**[92] (Hypothesis

Browser), εργαλείο για το σχεδιασμό υποθέσεων και την αξιολόγησή τους με συνέπεια ως προς τις υπάρχουσες γνώσεις, χρησιμοποιεί μια οντολογία υποθέσεων για να αναπαραστήσει υποθέσεις σε μορφή καταληπτή από μηχανή, ως σχέσεις μεταξύ αντικειμένων (πράκτορες) και διαδικασίες[93]. Ένα γεγονός υπόθεσης θεωρείται ως μια αφηρημένη βιολογική εκδήλωση. Η οντολογία φιλοξενεί διαθέσιμα στοιχεία βιβλιογραφίας, που προέρχονται κυρίως από τη βάση δεδομένων **Yeast Proteome Database**[94], σε πρώτο επίπεδο ανάλυσης. Το πρόγραμμα **Large-Scale Discovery of Scientific Hypotheses**[95] στοχεύει στο να συλλεχθούν και να γίνουν εμφανείς, συγκρίσιμες και υπολογίσιμα αντικρουόμενες (σε σχέση με ένα τυπικό παράδειγμα) υποθέσεις, οι οποίες παράγονται από κοινότητες που εστιάζονται σε τρεις κατηγορίες φαινότυπων ασθενειών (καρκίνο, νευροψυχιατρικές διαταραχές και λοιμώδη νοσήματα). Στο πρόγραμμα αυτό, οι υποθέσεις και τα στοιχεία που τις υποστηρίζουν συλλέγονται και δομούνται με τη μορφή δηλώσεων, ενώ στη συνέχεια επισημοποιούνται ως μια προτασιακή γραφική παράσταση.

Είναι πιθανό ότι η σημερινή πλειοψηφία των υποθέσεων στη βιολογία παράγεται από υπολογιστή. Οι υπολογιστές αυτοματοποιούν όλο και περισσότερο την διαδικασία διατύπωσης υποθέσεων όπως για παράδειγμα, διάφορα προγράμματα που χρησιμοποιούν εκμάθηση μηχανής (με βάση την επαγωγή) και εφαρμόζονται στη χημεία για τον σχεδιασμό φαρμάκων και στη βιολογία, όπου ο χαρακτηρισμός γονιδιωμάτων είναι ουσιαστικά μια μεγάλη διαδικασία αφαιρετικής διατύπωσης υποθέσεων. Τέτοιες υποθέσεις που παράγονται από υπολογιστή, έχουν απαραιτήτως διατυπωθεί σε υπολογιστικά δεκτή γλώσσα, αλλά δεν είναι ακόμα κοινή πρακτική να κατατεθούν σε μια δημόσια βάση δεδομένων και να καταστούν διαθέσιμες για επεξεργασία από άλλες εφαρμογές.

Κεφάλαιο 4: Διερεύνηση Ιατρικών Υποθέσεων για τον καρκίνο του τραχήλου της μήτρας

4.1 Εισαγωγή

Στα πλαίσια της παρούσας διπλωματικής εργασίας, θα γίνει μια διερεύνηση διαφόρων ιατρικών υποθέσεων που αφορούν τον καρκίνο του τραχήλου της μήτρας. Προτού προχωρήσουμε στην διερεύνηση με τη βοήθεια του λογισμικού **QDA Miner©**[96] και **WordStat©**[97] (Παράρτημα), παρουσιάζονται μερικά εισαγωγικά στοιχεία για τον καρκίνο του τραχήλου της μήτρας, τα οποία χρησιμοποιούνται για τον καθορισμό του πλαισίου, πάνω στο οποίο θα βασίσουμε την υλοποίηση της συγκεκριμένης διερεύνησης.

4.2 Επισκόπηση: Καρκίνος του τραχήλου της μήτρας

Ο καρκίνος του τραχήλου της μήτρας[98], όπως και άλλοι επιθηλιακοί όγκοι είναι πολύ σπάνιοι τις δυο πρώτες δεκαετίες της ζωής της γυναίκας, αποκτά όμως ιδιαίτερο ενδιαφέρον λόγω της ανησυχητικής αύξησης των προδιηθητικών μορφών του καρκίνου του τραχήλου της μήτρας σε νέες γυναίκες. Τόσο στην προδιηθητική όσο και στην διηθητική μορφή του, έχει αναγνωρισθεί ότι οφείλεται στην παρουσία του ιού των ανθρωπίνων θηλωμάτων (**Human Papilloma Virus: HPV**) στο τραχηλικό επιθήλιο.

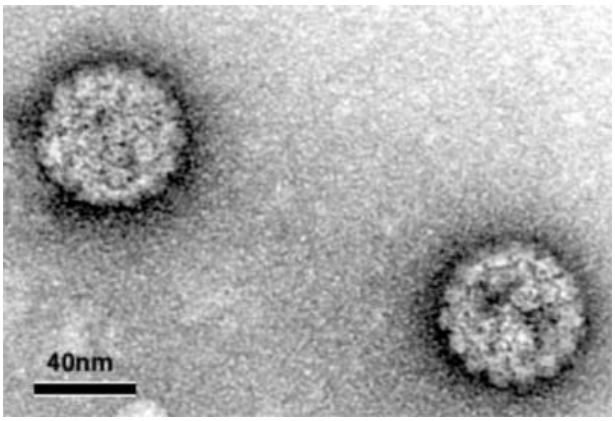
Εικόνα 10 *Ανατομία Γυναικείου Αναπαραγωγικού Συστήματος*

4.2.1 Αίτια

Ο καρκίνος του τραχήλου της μήτρας είναι μοναδικός στο ότι είναι ο πρώτος συμπαγής όγκος, ο οποίος έχει αποδειχθεί ότι οφείλεται αποκλειστικά σε ιό. Το συμπέρασμα αυτό στηρίζεται σε συγκλίνουσες αποδείξεις από διάφορα ερευνητικά κέντρα, που δείχνουν ότι ο ιός των ανθρωπίνων θηλωμάτων αποτελεί τον σπουδαιότερο παράγοντα στην καρκινογένεση του τραχηλικού επιθηλίου.

Οι ιοί HPV αποτελούν μια ετερόκλητη ομάδα DNA ιών, που σχετίζονται με υπερπλαστικές αλλοιώσεις (κονδυλώματα), δυσπλαστικές αλλοιώσεις (CIN) και κακοήθεις νεοπλασίες του πλακωδούς επιθηλίου του τραχήλου της μήτρας (καρκίνος). Οι ιοί HPV, είναι μικροί (55 nm), έχουν υψηλή εξειδίκευση για τον ξενιστή και καθένας από αυτούς είναι υπεύθυνος για την δημιουργία ειδικών ανατομικών και ιστολογικών αλλοιώσεων στον άνθρωπο. Μεταδίδοντα με την σεξουαλική επαφή, ανευρίσκονται συχνότερα σε νέες γυναίκες και προηγούνται της εμφάνισης ενδοεπιθηλιακών αλλοιώσεων κατά διάφορα χρονικά διαστήματα. Η γνώση της φυσικής ιστορίας της νόσου και ειδικότερα της σχέσης μεταξύ CIN και διηθητικού καρκίνου έχει μεγάλη σημασία, ιδιαίτερα στην αντιμετώπιση γυναικών με διαφορετικούς βαθμούς τραχηλικών ενδοεπιθηλιακών αλλοιώσεων.

Μελέτη της φυσικής ιστορίας του ιού, δείχνουν ότι στις περισσότερες περιπτώσεις η λοίμωξη είναι παροδική. Περίπου 60% έως 70% των νέων γυναικών θα είναι αρνητικές στην HPV λοίμωξη σε 30 περίπου μήνες από την εμφάνιση της αρχικής λοίμωξης[99]. Αντίθετα με την αυξημένη συχνότητα χαμηλού βαθμού ενδοεπιθηλιακών αλλοιώσεων σε νέες γυναίκες, η εμφάνιση υψηλού βαθμού ενδοεπιθηλιακών αλλοιώσεων δεν είναι συχνή (<7%) και οφείλεται στην παρουσία ιών υψηλού κινδύνου (HPV16, HPV18).



Εικόνα 11 Ιός HPV16

4.2.2 Παράγοντες Κινδύνου

Η έναρξη της σεξουαλικής δραστηριότητας σε μικρή ηλικία (<17 χρόνια) έχει αναγνωρισθεί από διάφορες επιδημιολογικές μελέτες σαν ο σπουδαιότερος παράγοντας κινδύνου στην ανάπτυξη προκαρκινικών αλλοιώσεων στον τράχηλο της μήτρας. Λοίμωξη του κατώτερου γεννητικού συστήματος σε νέες γυναίκες με τον ιό HPV, έχει αναδειχθεί σήμερα στο

συχνότερο σεξουαλικό μεταδιδόμενο νόσημα σε πολλές χώρες του κόσμου. Παθολογικές εξετάσεις κατά Παπανικολάου σε νέες γυναίκες, οφείλονται σχεδόν πάντα στην ύπαρξη HPV λοίμωξης και ιδιαίτερα των ιών της ομάδας χαμηλού κινδύνου.

Ο ιός συνήθως δεν εμφανίζει συμπτώματα, οπότε δεν είναι δυνατόν να γνωρίζει κάποιος αν φέρει τον ιό ή όχι. Για τις περισσότερες γυναίκες, ο ιός παραμένει σε λανθάνουσα φάση, ωστόσο, αν αυτό δεν πραγματοποιηθεί, υπάρχει μια πιθανότητα να προκαλέσει καρκίνο του τραχήλου με την πάροδο του χρόνου.

Άλλοι παράγοντες που μπορεί να αυξήσουν τον κίνδυνο εμφάνισης του καρκίνου του τραχήλου είναι:

- Το κάπνισμα.
- Εάν κάποιος είναι φορέας του ιού HIV ή βρίσκεται σε κάποια κατάσταση που προκαλεί ανοσοανεπάρκεια.
- Χρησιμοποίηση χαπιών ελέγχου γεννήσεων για μεγάλο χρονικό διάστημα (πέντε ή περισσότερα έτη).
- Γέννηση 3 η περισσότερων παιδιών
- Πολλαπλοί ερωτικοί σύντροφοι.

4.2.2 Συμπτώματα

Οι αλλαγές στα μη φυσιολογικά κύτταρα του τραχήλου της μήτρας από μόνες τους, σπάνια αποτελούν αιτία συμπτωμάτων. Η εμφάνιση συμπτωμάτων σχετίζεται με το αν αυτές οι αλλαγές των κυττάρων έχουν την τάση να εξελιχθούν σε καρκίνο του τραχήλου. Τα πιο χαρακτηριστικά συμπτώματα του καρκίνου του τραχήλου μπορεί να περιλαμβάνουν:

- Μη φυσιολογική αιμορραγία από τον κόλπο, ή ανεξήγητη αλλαγή στον εμμηνορροϊκό κύκλο.
- Αιμορραγία εξ’ επαφής, για παράδειγμα κατά τη διάρκεια συνουσίας ή κατά την τοποθέτηση διαφράγματος.
- Δυσπαρεύνια.
- Αιμορραγικό κολπικό έκκριμα.

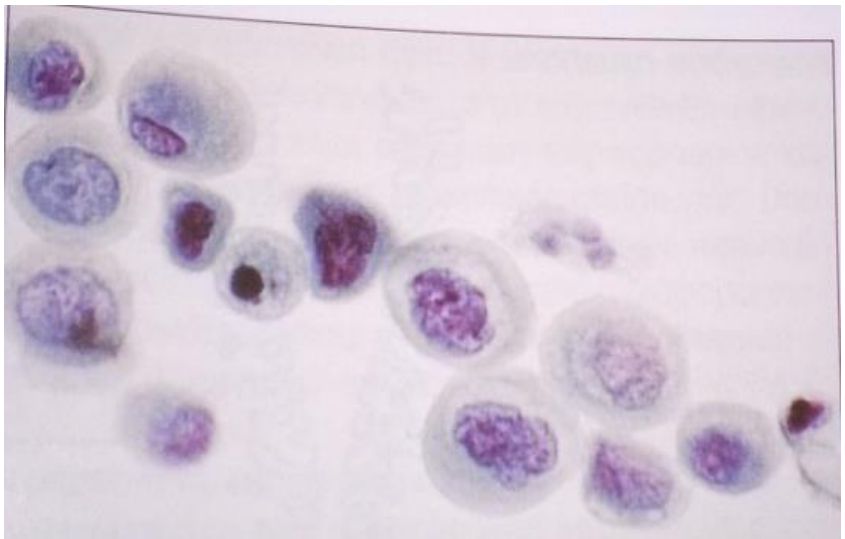
4.2.3 Διάγνωση

4.2.3.1 Κυτταρολογία (Screening)

Ο κυτταρολογικός έλεγχος αποτελεί σήμερα τον κύριο και μοναδικό ίσως τρόπο έγκαιρης διάγνωσης των πρόδρομων μορφών του καρκίνου του τραχήλου της μήτρας. Η εξέταση κατά Παπανικολάου (test Pap) χαρακτηρίστηκε σαν το μοναδικό ίσως αποτελεσματικό τρόπο μαζικού πληθυσμιακού ελέγχου "cancer screening test" που γνωρίζουμε σήμερα.

Η λήψη του κυτταρικού υλικού πρέπει να γίνεται με την κατάλληλη σπάτουλα από την επιφάνεια της ζώνης μετάπλασης και θα πρέπει να περιλαμβάνει το σημείο μετάπτωσης του κυλινδρικού επιθηλίου του ενδοτραχήλου στο πολύστοιβο πλακώδες επιθήλιο του εξωτραχήλου. Για το λόγο αυτό, ενδείκνυται η λήψη κυτταρικού υλικού και από τον ενδοτράχηλο με τη βοήθεια ειδικής βούρτσας. Η προσβολή του τραχηλικού επιθηλίου από τον ιό, προκαλεί ειδικές κυτταρολογικές αλλοιώσεις που χαρακτηρίζονται ως **κοιλοκυττάρωση** (koilocytosis). Τα κοιλοκύτταρα θεωρούνται παθογνωμικά και χαρακτηρίζουν την προσβολή του τραχηλικού επιθηλίου από τον ιό HPV.

Στην περίπτωση των τραχηλικών **ενδοεπιθηλιακών αλλοιώσεων**, τα κύτταρα του τραχηλικού επιθηλίου χαρακτηρίζονται από ανώμαλο πυρήνα, αύξηση της πυρηνικής/πρωτοπλασματικής αναλογίας και ανώριμη κυτταροπλασματική διαφοροποίηση (**Εικόνα 12**). Στην κυτταρολογία τα κύτταρα αυτά ονομάζονται δυσκαρυωτικά και ο κυτταρολογικός όρος δυσκαρύωση υποδιαιρείται σε ελαφρού, μετρίου και σοβαρού βαθμού ή σύμφωνα με την CIN ορολογία σε **τραχηλική ενδοεπιθηλιακή νεοπλασία**, ελαφρού, μετρίου ή σοβαρού βαθμού (CIN1, CIN2 και CIN3)



Εικόνα 12 *Υψηλού βαθμού ενδοεπιθηλιακή αλλοίωση σε κυτταρολογικό παρασκεύασμα*

Οι περισσότερες πρόσφατες εργασίες προτείνουν τον έλεγχο με εξέταση κατά Παπανικολάου όλων των νέων κοριτσιών που έχουν οποιαδήποτε σεξουαλική δραστηριότητα ανεξαρτήτως ηλικίας.

4.2.3.2 Κολποσκοπική εξέταση

Η κολποσκόπηση μπορεί να καθορίσει την τοπογραφία και την έκταση της τραχηλικής βλάβης και να επιβεβαιώσει ιστολογικά τον βαθμό της τραχηλικής εξεργασίας, στοιχείο απαραίτητο για τη σωστή επιλογή κατάλληλης θεραπευτικής αντιμετώπισης των ασθενών. Το κολποσκόπιο είναι ένα διοπτρικό μικροσκόπιο, με το οποίο είναι δυνατόν να εξεταστεί το τραχηλικό επιθήλιο και η αγγείωση του, σε μεγέθυνση τάξης από 6 έως 40 φορές. Η κολποσκοπική εκτίμηση στηρίζεται στην αξιολόγηση τριών βασικών χαρακτηριστικών του επιθηλίου, που είναι τα εξής: το χρώμα, η μορφολογία της επιφάνειας και η αρχιτεκτονική των αγγείων. Η κολποσκόπηση είναι απαραίτητη για τη λήψεις κατευθυνόμενων βιοψιών από τις ύποπτες περιοχές του τραχήλου, οι οποίες και τελικά θα καθορίσουν ιστολογικά τον βαθμό της επιθηλιακής βλάβης.

Η κυτταρολογία και η κολποσκόπηση, είναι δυο μέθοδοι που αλληλοσυμπληρώνονται και ο συνδυασμός των δυο παίζει ένα σπουδαίο ρόλο στον έλεγχο και αντιμετώπιση των ασθενών με τραχηλικές ενδοεπιθηλιακές αλλοιώσεις. Όλες οι γυναίκες με παθολογική εξέταση κατά Παπανικολάου, θα πρέπει να ελέγχονται κολποσκοπικά πριν από την επιλογή οποιασδήποτε θεραπευτικής αντιμετώπισης, γιατί μόνο με αυτόν τον τρόπο μπορεί να εκτιμηθεί σωστά η εντόπιση, η έκταση και ο βαθμός της ενδοεπιθηλιακής βλάβης, που θα επιτρέψει την επιλογή της σωστής θεραπευτικής αντιμετώπισης.

4.2.4 Σταδιοποίηση του καρκίνου του τραχήλου της μήτρας κατά FIGO[100]

Προδιηθητικό Καρκίνωμα			
Στάδιο 0			Καρκίνος in situ, ενδοεπιθηλιακός καρκίνος (οι περιπτώσεις του σταδίου 0 δεν πρέπει να περιλαμβάνονται σε καμία θεραπευτική στατιστική)
Διηθητικό Καρκίνωμα			
Στάδιο I			Ο καρκίνος περιορίζεται αυστηρά στον τράχηλο (δεν πρέπει να λαμβάνεται υπόψη η έκταση στο σώμα της μήτρας)
	Στάδιο Ia		Προκλινικός Καρκίνος του τραχήλου που διαγιγνώσκεται μόνο με μικροσκόπηση
		Στάδιο Ia1	Εστίες με βάθος διήθησης 3mm
		Στάδιο Ia2	Εστίες που ανιχνευονται μικροσκοπικά και μπορούν να υπολογιστούν οι διαστάσεις τους. Το βάθος διήθησης είναι > 3-5 mm υπολογιζόμενο από τη βάση του επιθηλίου, είτε του επιφανειακού είτε του αδενικού, από το οποίο προέρχεται και η δεύτερη διάσταση, η οριζόντια επέκταση δεν πρέπει να ξεπερνά τα 7 mm. Μεγαλύτερες βλάβες πρέπει να κατατάσσονται στο στάδιο 1b.
	Στάδιο Ib		Διηθητικές βλάβες > 5mm
		Στάδιο Ib1	Βλάβη με βάθος μικρότερο από ή ίση με 4 cm
		Στάδιο Ib2	Βλάβες με βάθος πάνω από 4 cm
Στάδιο II			Ο καρκίνος επεκτείνεται πέρα από τον τράχηλο αλλά δεν έχει φθάσει στα πυελικά τοιχώματα. Έχει καταλάβει και τον κόλπο αλλά όχι το κάτω τριτομόριο.
	Στάδιο IIa		Χωρίς εμφανή διήθηση των παραμητριών
	Στάδιο IIb		Εμφανή διήθηση των παραμητρίων
Στάδιο III			Ο καρκίνος έχει επεκταθεί στα πυελικά τοιχώματα. Στην εξέταση από το ορθό δεν υπάρχει ελεύθερο διάστημα από καρκίνο μεταξύ του όγκου και του πυελικού τοιχώματος. Ο όγκος καταλαμβάνει και το κατώτερο τριτημόριο του κόλπου. Όλες οι περιπτώσεις με υδρονέφρωση ή μη λειτουργούντα νεφρά
	Στάδιο IIIa		Χωρίς επέκταση στο πυελικό τοίχωμα
	Στάδιο IIIb		Επέκταση στο πυελικό τοίχωμα με/ή υδρονέφρωση ή μη λειτουργικό νεφρό.
Στάδιο IV			Ο καρκίνος έχει επεκταθεί πέρα από την ελάσσονα πύελο ή έχει κλινικά καταλάβει το βλεννογόνο της κύστης ή του ορθού. Ένα φυσαλιδώδες οίδημα μόνο δεν επιτρέπει μια περίπτωση να συμπεριληφθεί στο στάδιο IV
	Στάδιο IVa		Επέκταση στα γειτονικά όργανα
	Στάδιο IVb		Επέκταση στα απομακρυσμένα όργανα.

4.2.5 Θεραπεία

4.2.5.1 Αντιμετώπιση των χαμηλού βαθμού τραχηλικών ενδοεπιθηλιακών αλλοιώσεων (LGSIL)

Οι περισσότερες νέες γυναίκες με LGSIL μπορούν να αντιμετωπιστούν με παρακολούθηση για συγκεκριμένο χρονικό διάστημα, συνήθως ενός η δυο χρόνων. Η παρακολούθηση αυτή μπορεί να γίνει κυτταρολογικά, κολποσκοπικά και με HPV-DNA ανάλυση. Η τεχνολογία της ανίχνευση του DNA του ιού, έχει εξελιχθεί τα τελευταία χρόνια και οι σπουδαιότερες τεχνικές που χρησιμοποιούνται είναι:

i. **Αλυσωτή Αντίδραση Πολυμεράσης** (**Polymerase Chain Reaction: PCR**). Η εξέταση αυτή έχει μεγάλη ευαισθησία για την ανίχνευση πολύ χαμηλών συγκεντρώσεων μορίων DNA του ιού.
ii. **Κυρίαρχο Σύστημα Υβριδισμού** (**Hybrid Capture System: HCS**). Η μεθοδολογία αυτή είναι ένα σύστημα υβριδισμού υγρής φάσης για τον ποσοτικό προσδιορισμό μεγαλύτερων συγκεντρώσεων HPV DNA.

4.2.5.2 Αντιμετώπιση των υψηλού βαθμού τραχηλικών ενδοεπιθηλιακών αλλοιώσεων (HGSIL)

Η επιμονή της τραχηλικής βλάβης για μακρό χρονικό διάστημα και η ανίχνευση ιών υψηλού κινδύνου αποτελούν τις συχνότερες ενδείξεις θεραπείας. Η βασική αρχή της μοντέρνας θεραπείας όλων των προκαρκινικών αλλοιώσεων του τραχήλου της μήτρας είναι η ακριβής εντόπιση της παθολογικής περιοχής του επιθηλίου και αυτό είναι δυνατό μόνο με τη βοήθεια του κολποσκοπίου. Με την πάροδο της τεχνολογίας, το ενδιαφέρον επικεντρώνεται σε νεότερες μεθόδους τοπικής καταστροφής ή αφαίρεσης της παθολογικής περιοχής και αυτοί οι τρόποι θεραπείας, αποκτούν μεγαλύτερη σημασία παρά στο παρελθόν, λόγω της ανάγκης διατήρησης της αναπαραγωγικής ικανότητας των νέων γυναικών.

Παρακάτω αναφέρονται ονομαστικά οι κυριότερες μέθοδοι αντιμετώπισης των HGSIL:

- Αγκύλη Διαθερμίας Leep-Lietz
- LASER CO_2
- Θερμική Εξάχνωση
- Κωνοειδής εκτομή με LASER CO_2
- Κρυοθεραπεία
- Ηλεκτρικός Καυτηριασμός

4.2.5.3 Αντιμετώπιση Διηθητικού Καρκίνου

Οι δυο βασικοί τρόποι θεραπείας του καρκίνου του τραχήλου της μήτρας είναι η χειρουργική θεραπεία και η ακτινική. Η χειρουργική θεραπεία συνίσταται σε υστερεκτομία μετά ή άνευ των εξαρτημάτων. Η επέμβαση περιλαμβάνει εκτεταμένη πυελική λεμφαδενεκτομία, την αφαίρεση του μεγαλύτερου τμήματος των κυρίων συνδέσμων και την ιερομητρικών συνδέσμων της μήτρας, καθώς και του άνω τριτημορίου του κόλπου. Η

ακτινική θεραπεία μπορεί να εφαρμοσθεί σε όλα τα στάδια του καρκίνου του τραχήλου της μήτρας, ενώ η χειρουργική εφαρμόζεται μόνο στα στάδια I και IIa.

4.3 Παρουσίαση Πλαισίου Αναφοράς

Σύμφωνα με την προηγούμενη παράγραφο, καθώς και από άλλες ιατρικές πηγές, οι πιο κοινοί όροι που σχετίζονται με τον καρκίνο του τραχήλου της μήτρας μπορούν να παρουσιαστούν υπό τη μορφή της παρακάτω λίστας όρων – keywords **{L}** :

{L} = {cervical carcinoma, human papillomavirus, HPV, test PAP, LEEP, hysterectomy, E6 oncoprotein, E7 oncoprotein},

όπου:

- **Cervical carcinoma** : ο όρος αυτός είναι πιο ειδικός για τον καρκίνο στο επιθήλιο του τραχήλου της μήτρας, εκεί που δρα δηλαδή ο HPV
- **Human Papillomavirus** ή **HPV** : ο ιός ανθρωπίνων θηλωμάτων
- **Test PAP** : η εξέταση κατά Παπανικολάου
- **LEEP** : αρχικά του Loop Electrosurgical Excision Procedure (ηλεκτροχειρουργική διαδικασία εκτομής με αγκύλη) που αποτελεί μια μέθοδο εκτομής της αλλοιωμένης περιοχής του τραχήλου
- **Hysterectomy** : η χειρουργική επέμβαση της υστερεκτομής.
- **E6/E7 oncoproteins** : οι ογκοπρωτεΐνες E6 και E7 είναι πρωτεΐνες που παράγει ο ιός HPV και σχετίζονται με υψηλή πιθανότητα εμφάνισης καρκίνου του τραχήλου της μήτρας

Οι παραπάνω όροι μπορούν να αποτελέσουν ένα αρχικό πλαίσιο όρων **A**, **C** σύμφωνα με το μοντέλο υποθέσεων του Swanson, όπως αναφέρθηκε στην παράγραφο 3.3.1, ώστε να ελέγξουμε την πιστότητα του μοντέλου. Αρχικά θα εφαρμόσουμε το μοντέλο με δυο από τους παραπάνω όρους, που εκ των προτέρων γνωρίζουμε ότι αλληλοσυνδέονται. Επιλέγοντας τους λιγότερο γνωστούς όρους, ορίζουμε ως:

- **Όρος A: Loop Electrosurgical Excision**
- **Όρος C: E6 or E7 oncoprotein**

Στην ιατρική υπόθεση που διατυπώνουμε, θεωρούμε ότι οι παραπάνω όροι ανήκουν σε ασύνδετες φαινομενικά βιβλιογραφίες και θέλουμε να ερευνήσουμε την ύπαρξη ή έλλειψη κοινής ορολογίας ανάμεσα στις βιβλιογραφίες, ψάχνουμε δηλαδή πιθανούς όρους **B**.

4.3.1 Μεθοδολογία

Έχοντας ορίσει τους όρους **A**, **C**, σύμφωνα με το μοντέλο υποθέσεων του Swanson, πρέπει να ανατρέξουμε στις αντίστοιχες βιβλιογραφίες. Για να εφαρμόσουμε την παραπάνω λογική, αρχικά θα εντοπίσουμε τις βιβλιογραφίες των όρων **A**, **C** (οι οποίες στα πλαίσια της παρούσας διπλωματικής, αποτελούνται από άρθρα, αποτελέσματα ερευνών και δημοσιεύσεων που βρίσκονται στο **PubMed©**) και στη συνέχεια θα εφαρμόσουμε τεχνικές εξόρυξης κειμένου ώστε να βρούμε όρους **B**, που μπορεί να οδηγούν σε σύνδεση μεταξύ των δυο φαινομενικά ασύνδετων βιβλιογραφιών.

Η εφαρμογή της μεθόδου απλοποιείται ακόμα περισσότερο, θεωρώντας τα παρακάτω βήματα:

1. Από το τεράστιο πλήθος πληροφοριών που περιέχονται στην εκάστοτε βιβλιογραφία, αυτό που μας ενδιαφέρει κυρίως είναι η περίληψη (**Abstract Text**) διότι αφενός μεν περιλαμβάνονται σημαντικές πληροφορίες όπως οι λέξεις-κλειδιά (keywords), αφετέρου όμως δεν υπάρχει παραπλανητική πληροφορία όπως αυτή μπορεί να εμφανίζεται στους τίτλους των άρθρων.

2. Αποκτώντας με κάποια τεχνική εξόρυξης μόνο το κείμενο των περιλήψεων από την βιβλιογραφία, μπορούμε να υπολογίσουμε το πλήθος των λέξεων που εμφανίζονται συχνότερα στη συγκεκριμένη βιβλιογραφία, καθώς και ποιες είναι αυτές, χωρίς να χρειαστεί να προχωρήσουμε σε πολύπλοκες διαδικασίες αναγνώρισης οντοτήτων και τεχνικών χαρακτηριστικών από τα κείμενα. Πρακτικά, χρειαζόμαστε μόνο τον κατακερματισμό του κειμένου και ένα μέτρο σχετικότητας.

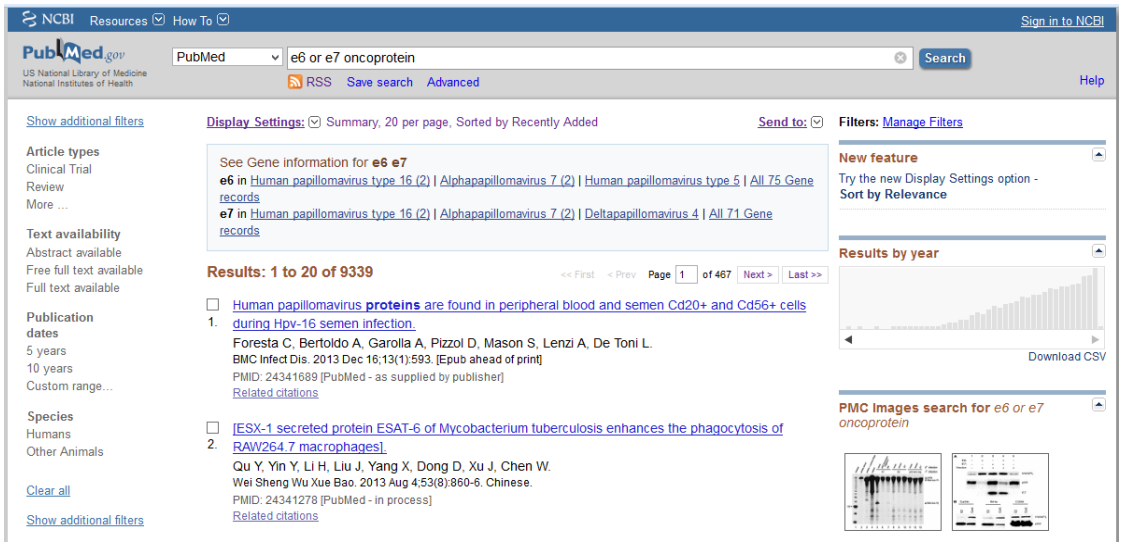
3. Αφού υλοποιήσουμε τα παραπάνω βήματα και για τις δυο βιβλιογραφίες, μπορούμε να προχωρήσουμε σε σύγκριση των δύο συνόλων με τους πιο συνήθεις όρους στα κείμενα των δυο βιβλιογραφιών και να επαληθεύσουμε την ύπαρξη ή την έλλειψη κοινών όρων, που μπορεί να οδηγούν σε σύνδεση των δυο βιβλιογραφιών.

4.3.2 Εφαρμογή Μεθοδολογίας

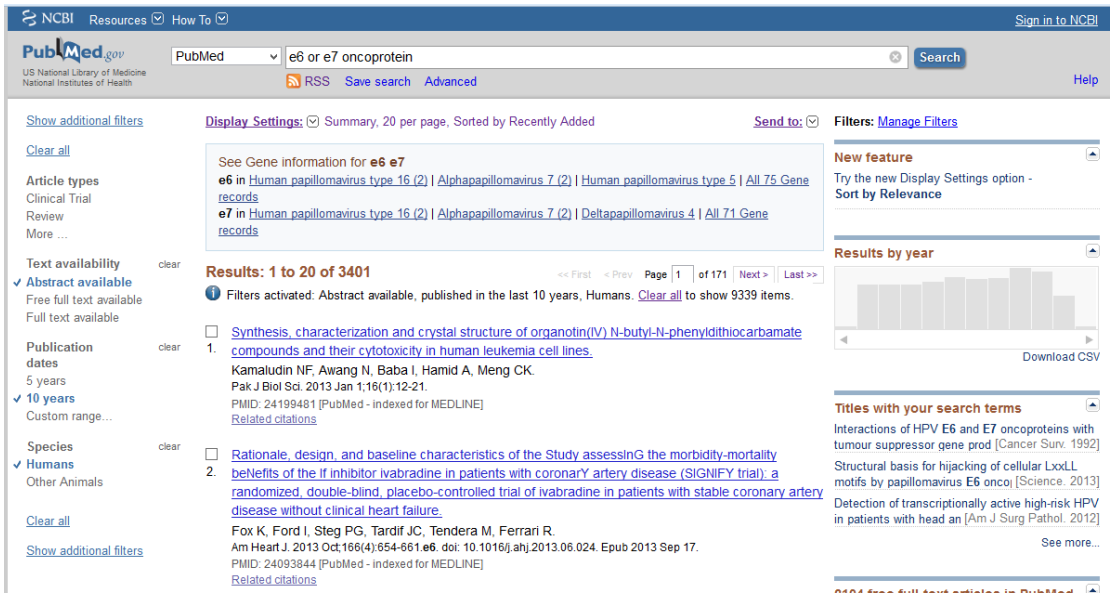
Αρχικά μέσω ενός browser, κατευθυνόμαστε στο site του **PubMed©** (http://www.ncbi.nlm.nih.gov/pubmed) και στο πεδίο αναζήτησης γράφουμε τον όρο **C**: **E6 or E7 oncoprotein**. Η αναζήτηση μας επιστρέφει περί τα 9400 άρθρα που σχετίζονται με τον συγκεκριμένο όρο (**Εικόνα 13**). Τα φίλτρα που βρίσκονται στην αριστερή πλευρά της ιστοσελίδας, μας δίνουν τη δυνατότητα να περιορίσουμε τα αποτελέσματα, οπότε επιλέγουμε τις εξής παραμέτρους:

- **Text Availability**: Abstract available
- **Publication dates**: 10 years
- **Species**: Human

Η ημερομηνία έκδοσης τέθηκε στα 10 χρόνια, διότι μέχρι το 2003 δεν υπήρχε ικανοποιητική ποσότητα άρθρων και δημοσιεύσεων που να έχει δημοσιευθεί πάνω στο θέμα του καρκίνου του τραχήλου της μήτρας και των παραπλήσιων προς αυτό όρους. Βάσει των επιλογών που κάναμε, παίρνουμε ένα σύνολο 3401 άρθρων, όπως φαίνεται στην **Εικόνα 14**.

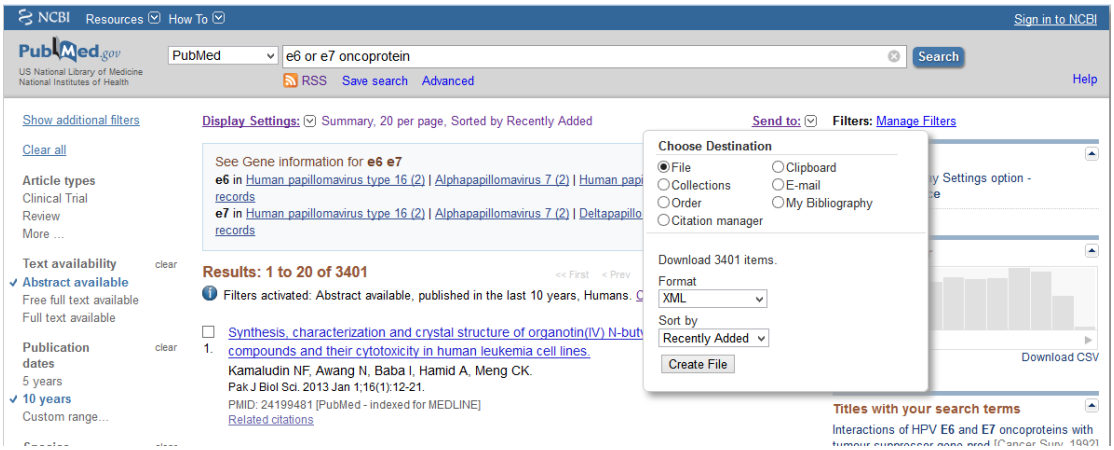


Εικόνα 13 Αποτελέσματα αναζήτησης στο Pubmed©, για τον όρο "E6 or E7 oncoprotein"



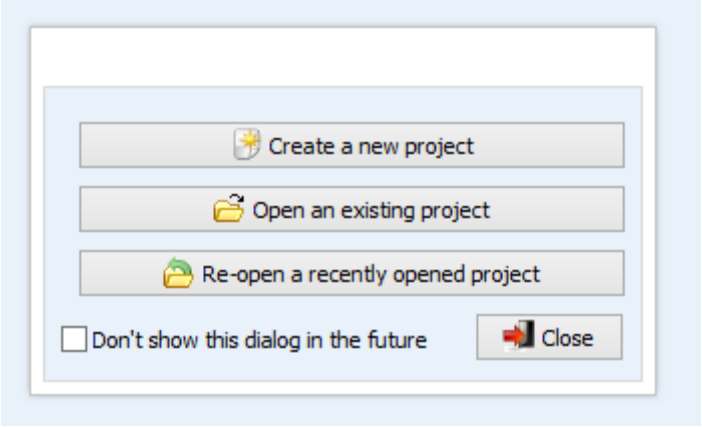
Εικόνα 14 Εφαρμογή φίλτρων στα αποτελέσματα αναζήτησης

Επιλέγουμε να αποθηκεύσουμε το αρχείο σε μορφή .XML, με την επιλογή *Send to*, όπως φαίνεται στην **Εικόνα 15** και αφού ολοκληρωθεί η αποθήκευση του αρχείου, χρησιμοποιούμε το λογισμικό ανάλυσης δεδομένων **QDA Miner**© για να επεξεργαστούμε τα δεδομένα.



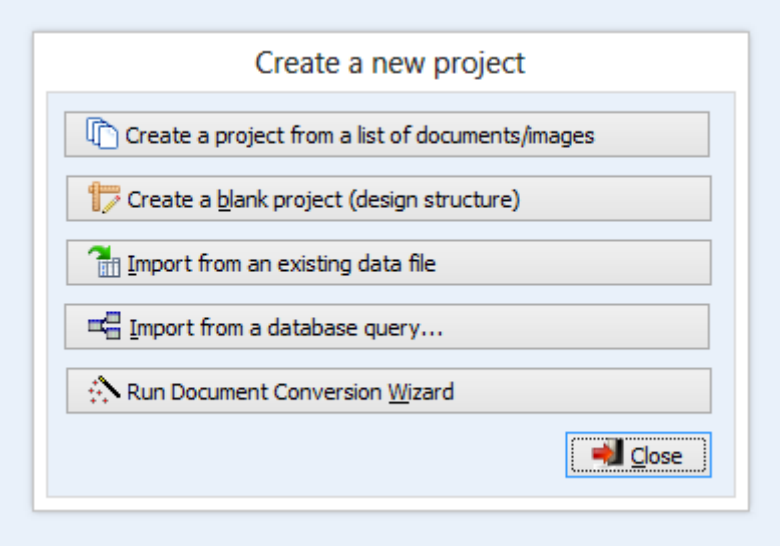
Εικόνα 15 Αποθήκευση αρχείου αποτελεσμάτων σε μορφή XML

Στην αρχική οθόνη επιλέγουμε Create a new project (**Εικόνα 16**) και στο αναδυόμενο παράθυρο επιλέγουμε το Import from an existing data file (**Εικόνα 17**), ώστε να εισάγουμε το αρχείο δεδομένων που έχουμε αποθηκεύσει από το **PubMed©**.

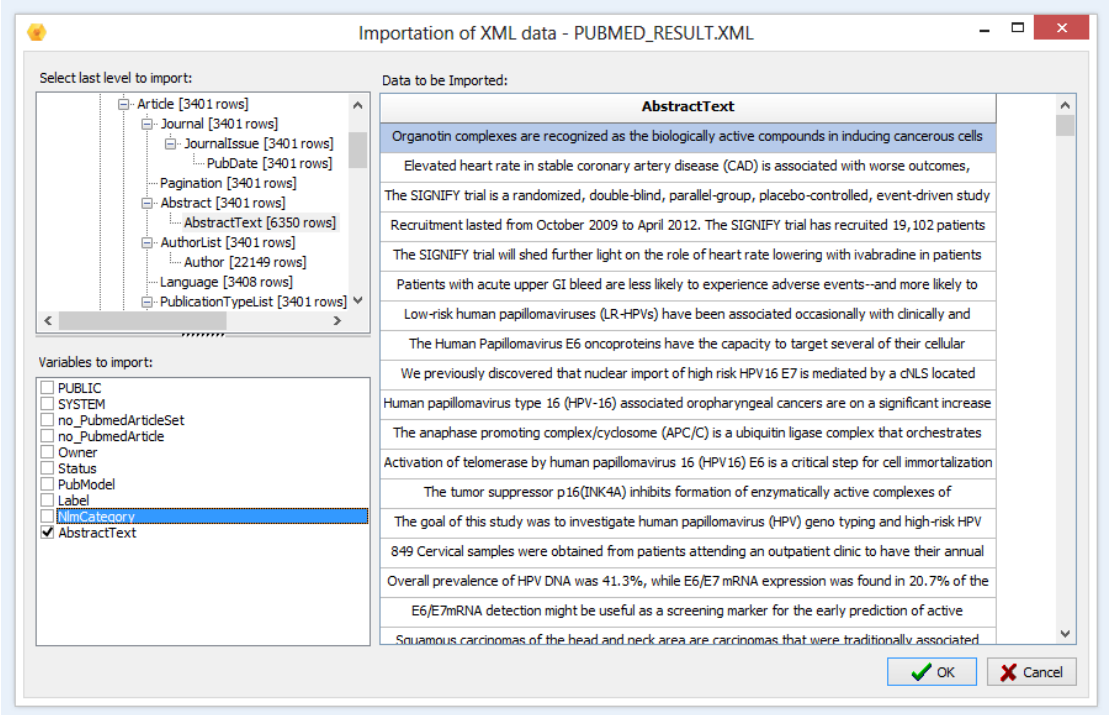


Εικόνα 16 Αρχική οθόνη προγράμματος QDA Miner©

Αφού μας δοθεί η επιλογή να ονομάσουμε το project και να το αποθηκεύσουμε σε κάποιο χώρο στο σκληρό δίσκο, εμφανίζεται η οθόνη παραμετροποίησης του αρχείου δεδομένων, από την οποία επιλέγουμε ποιες πληροφορίες θέλουμε να εισάγουμε για επεξεργασία στο πρόγραμμα (**Εικόνα 18).** Στη συγκεκριμένη περίπτωση, όπως αναφέραμε στην ενότητα 4.3.1, μας ενδιαφέρει μόνο το **Abstract Text**, οπότε και επιλέγουμε το αντίστοιχο πλαίσιο.



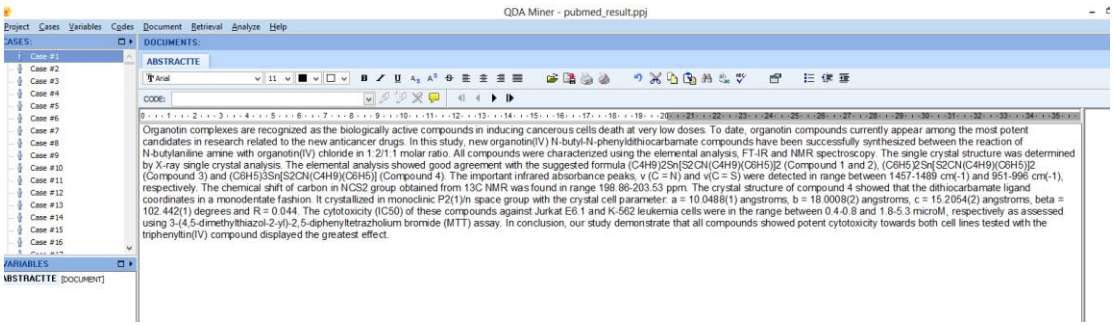
Εικόνα 17 Επιλογή αρχείου δεδομένων



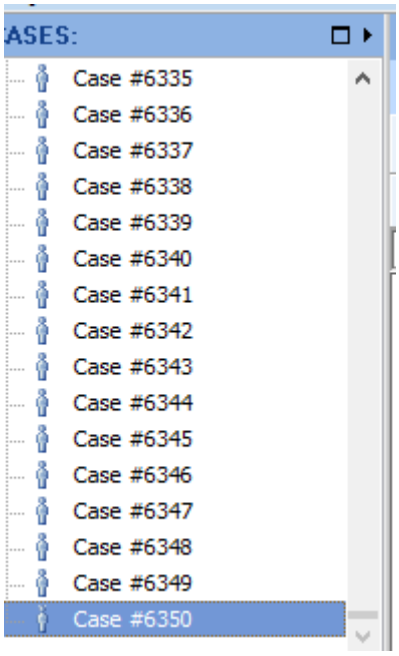
Εικόνα 18 Επιλογή πληροφοριών προς επεξεργασία

Πατώντας OK και μετά από κάποια χρονική διάρκεια, η οποία εξαρτάται από το πλήθος των δεδομένων που έχει να διαχειριστεί το πρόγραμμα, παίρνουμε τα αποτελέσματα (**Εικόνα 19**), στα οποία διακρίνουμε ότι στην κάθε περίπτωση (**Case #**) έχει αντιστοιχιστεί μόνο το κείμενο της περίληψης. Παρατηρούμε επιπλέον, ότι το πλήθος των περιπτώσεων είναι 6350 (**Εικόνα 20**), αριθμός που είναι σχεδόν διπλάσιος του αριθμού των άρθρων που περιέχει η συγκεκριμένη βιβλιογραφία. Η εξήγηση για τον αριθμό αυτό γίνεται εύκολα

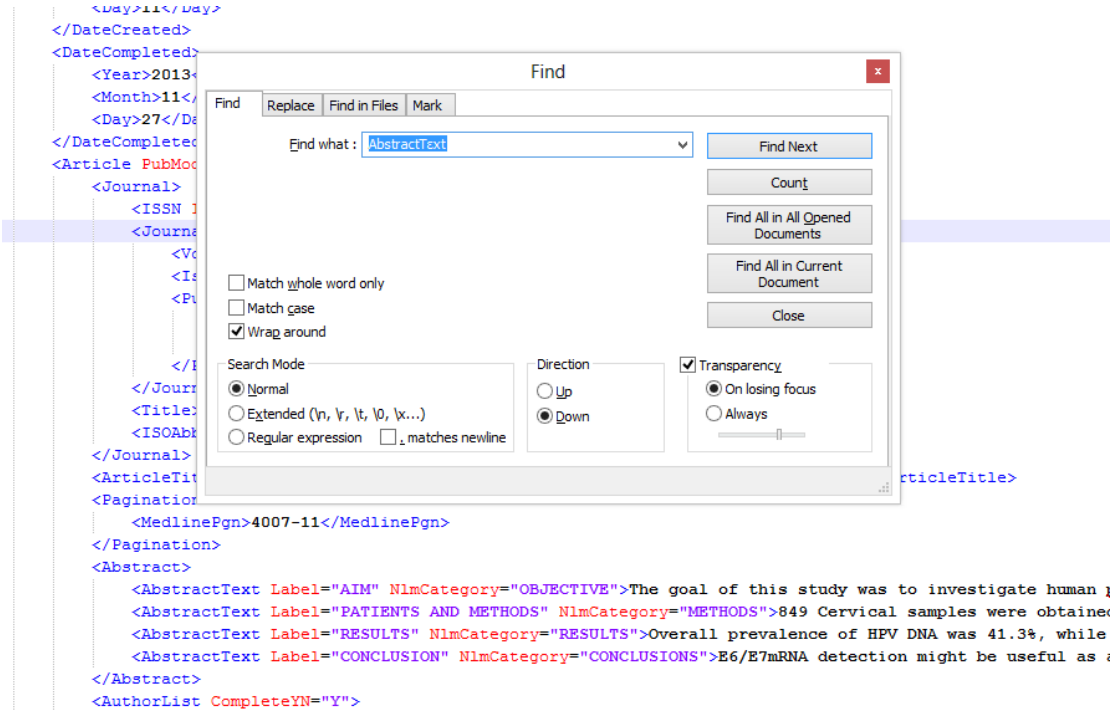
αντιληπτή, όταν ανοίξουμε το αρχείο .XML που περιέχει τη βιβλιογραφία: εφαρμόζοντας μια αναζήτηση του όρου **Abstract Text**, παρατηρούμε ότι σε κάποια άρθρα, κυρίως σε όσα προέρχονται από δημοσίευση αποτελεσμάτων κλινικών ερευνών, περιέχονται περιλήψεις για τα διάφορα στάδια της έρευνας (**Εικόνα 21**).



Εικόνα 19 Αποτελέσματα Επεξεργασίας Δεδομένων



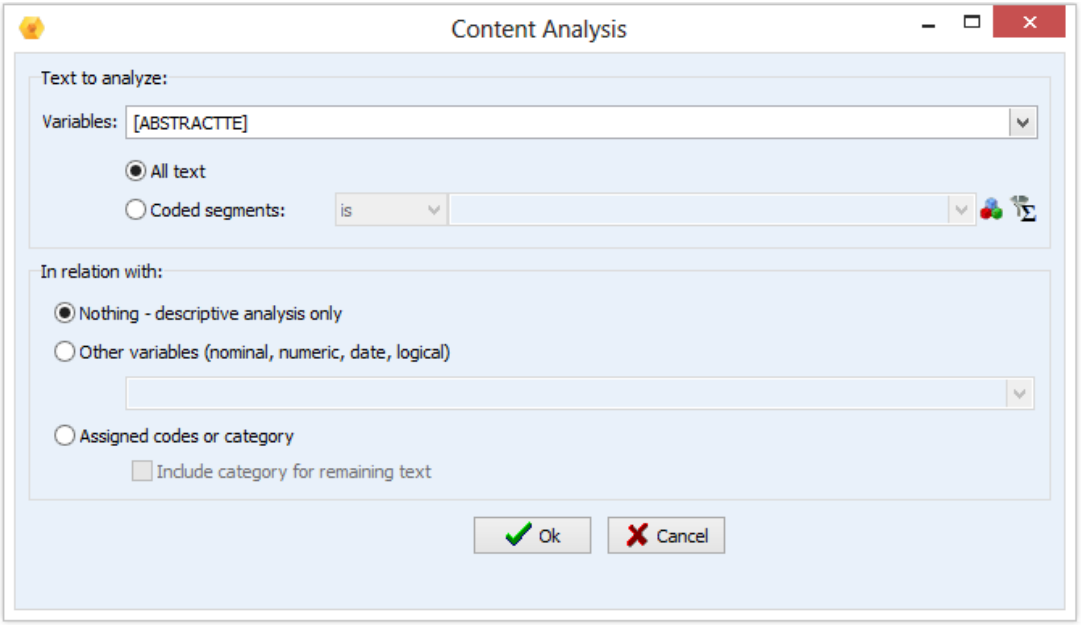
Εικόνα 20 Πλήθος Περιπτώσεων μετά από την Επεξεργασία Δεδομένων



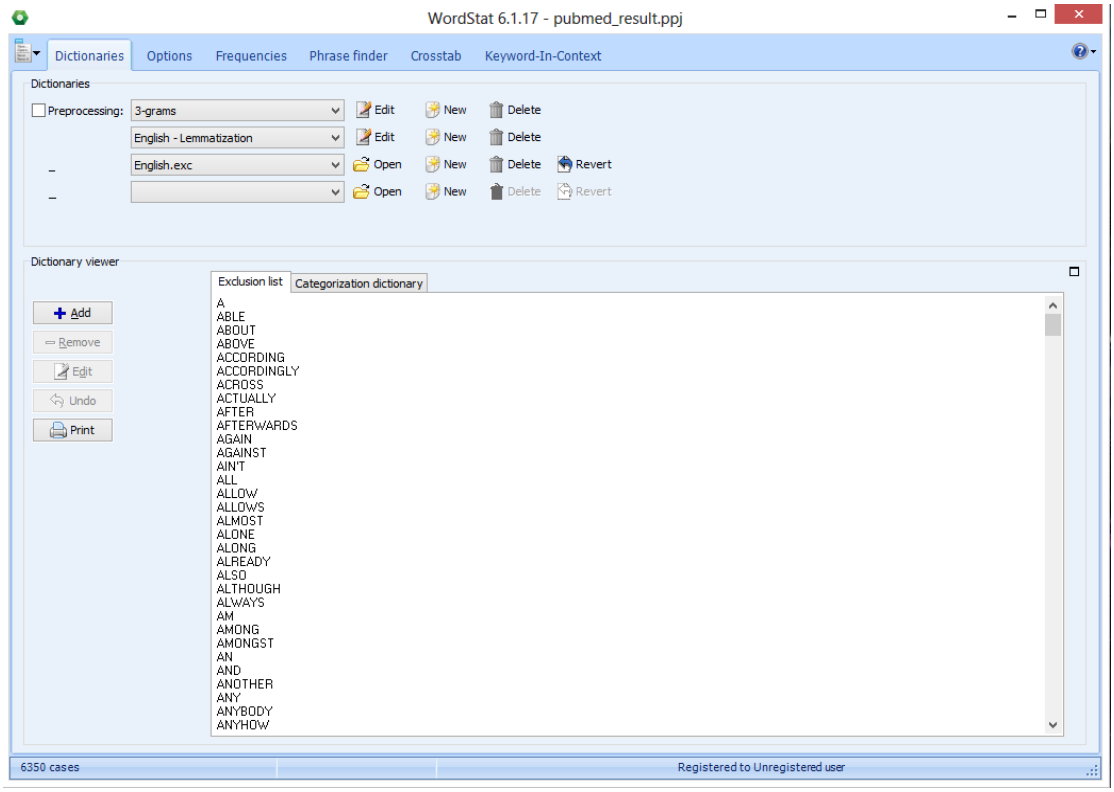
Εικόνα 21 Η διαφορά στο πλήθος των περιλήψεων, οφείλεται στην ύπαρξη πολλαπλών περιλήψεων, κυρίως σε άρθρα που προέρχονται από δημοσιεύσεις κλινικών αποτελεσμάτων από διάφορες έρευνες

Το επόμενο βήμα περιλαμβάνει την επεξεργασία των δεδομένων, με σκοπό την εξαγωγή χρήσιμης πληροφορίας. Από τη γραμμή εργασιών του **QDA Miner©**, επιλέγουμε Analyze → Context Analysis, καθώς και τα δεδομένα που θέλουμε να υποστούν επεξεργασία (ABSTRACTTE), όπως φαίνεται στην **Εικόνα 22**. Τα δεδομένα αυτά εισάγονται στην πλατφόρμα λογισμικού **WordStat©**, στην οποία μας δίνεται ήδη μια λίστα από **stopwords** (**Exclusion List**) μαζί με διάφορες άλλες επιλογές (**Εικόνα 23**). Στην καρτέλα Options, επιλέγουμε το φίλτρο συχνοτήτων που θέλουμε να εφαρμόσουμε στα δεδομένα, εν προκειμένω με το μέτρο **TF*IDF**, όπως έχει οριστεί στην ενότητα 1.3.3.1, ώστε να δούμε επιπλέον, πόσο σημαντικοί είναι οι όροι που εμφανίζονται συχνότερα (**Εικόνα 24**) και τελικά παίρνουμε τα αποτελέσματα όπως φαίνονται στην **Εικόνα 25**.

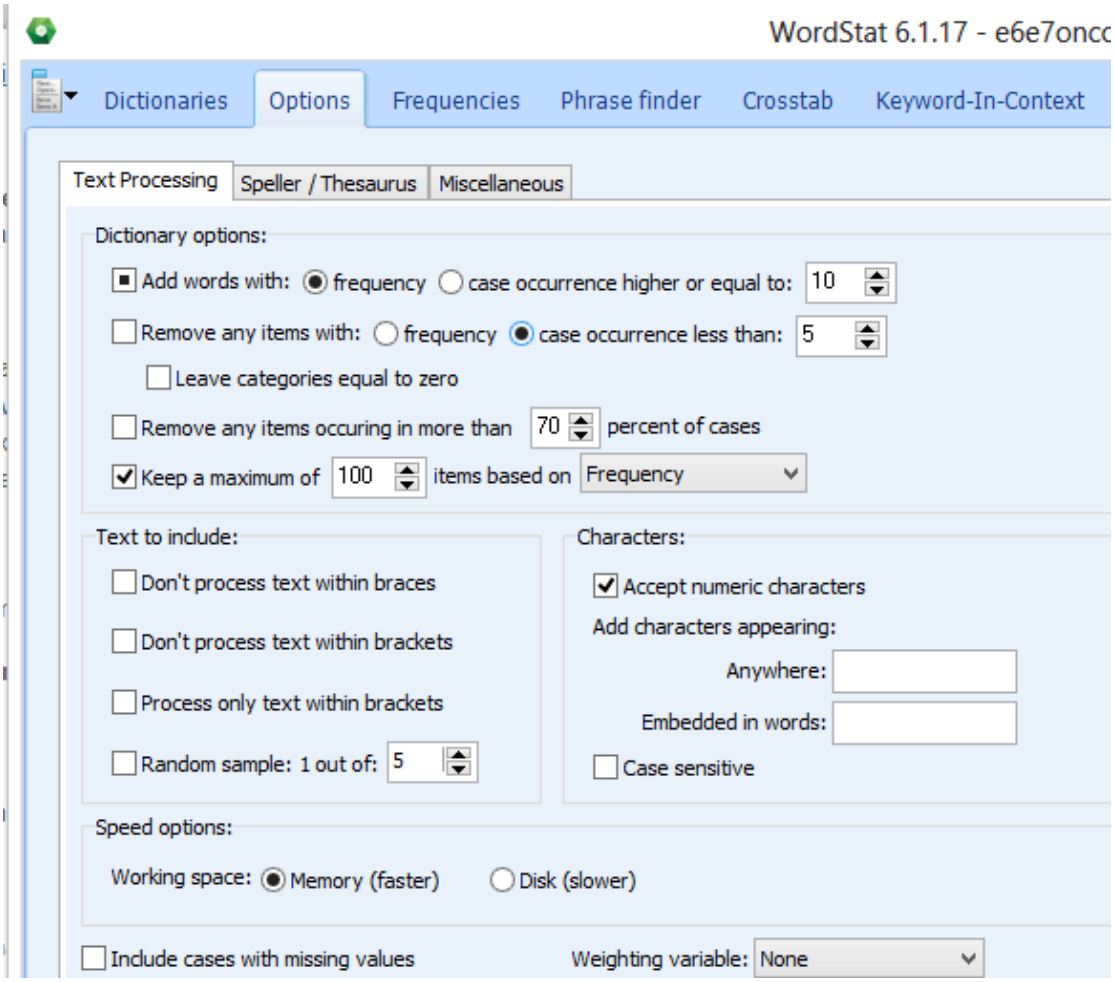
Σύμφωνα με τα αποτελέσματα αυτά, εξάγουμε τον **Πίνακα 6** για την βιβλιογραφία **C**, που αντιστοιχεί στον όρο ***E6 or E7 oncoprotein***. Όπως παρατηρούμε, κάποιοι όροι δεν μας δίνουν ιατρική πληροφορία (**patients**, **high**, **study**, **risk**, **positive**) και υπάρχει και επανάληψη όρων (**cells-cell**), οπότε εισάγουμε αυτές τις λέξεις στο **Exclusion List**, και επαναλαμβάνουμε τις μετρήσεις. Τα αποτελέσματα απεικονίζονται στον **Πίνακα 7** και η γραφική τους αναπαράσταση στην **Εικόνα 26**.



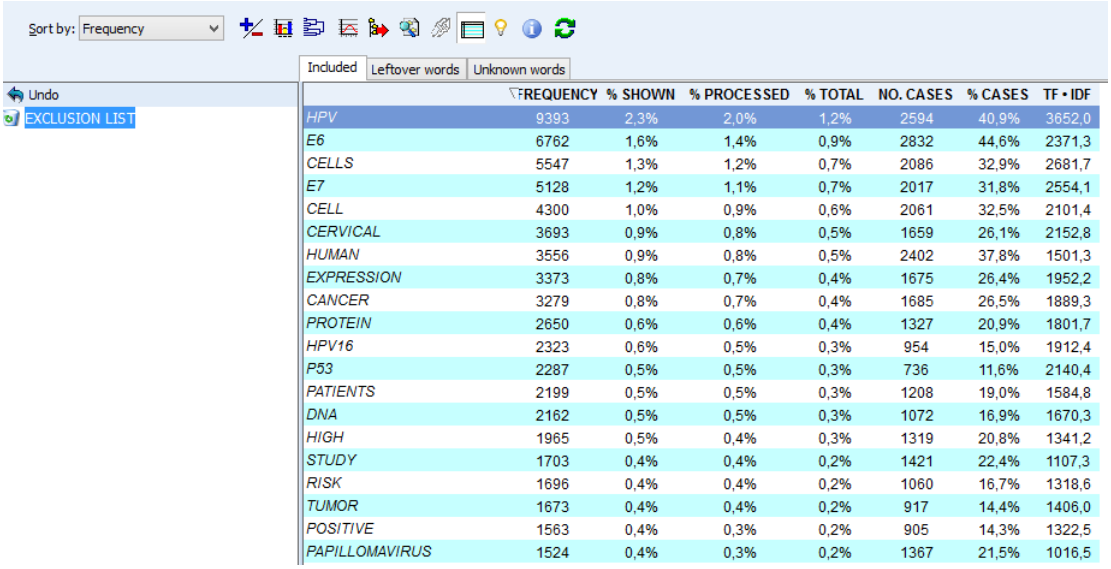
Εικόνα 22 Επιλογή Δεδομένων για ανάλυση



Εικόνα 23 Περιβάλλον WordStat©



Εικόνα 24 Εφαρμογή Επιλογών στον Τρόπο εμφάνισης Δεδομένων



	FREQUENCY	% SHOWN	% PROCESSED	% TOTAL	NO. CASES	% CASES	TF • IDF
HPV	9393	2,3%	2,0%	1,2%	2594	40,9%	3652,0
E6	6762	1,6%	1,4%	0,9%	2832	44,6%	2371,3
CELLS	5547	1,3%	1,2%	0,7%	2086	32,9%	2681,7
E7	5128	1,2%	1,1%	0,7%	2017	31,8%	2554,1
CELL	4300	1,0%	0,9%	0,6%	2061	32,5%	2101,4
CERVICAL	3693	0,9%	0,8%	0,5%	1659	26,1%	2152,8
HUMAN	3556	0,9%	0,8%	0,5%	2402	37,8%	1501,3
EXPRESSION	3373	0,8%	0,7%	0,4%	1675	26,4%	1952,2
CANCER	3279	0,8%	0,7%	0,4%	1685	26,5%	1889,3
PROTEIN	2650	0,6%	0,6%	0,4%	1327	20,9%	1801,7
HPV16	2323	0,6%	0,5%	0,3%	954	15,0%	1912,4
P53	2287	0,5%	0,5%	0,3%	736	11,6%	2140,4
PATIENTS	2199	0,5%	0,5%	0,3%	1208	19,0%	1584,8
DNA	2162	0,5%	0,5%	0,3%	1072	16,9%	1670,3
HIGH	1965	0,5%	0,4%	0,3%	1319	20,8%	1341,2
STUDY	1703	0,4%	0,4%	0,2%	1421	22,4%	1107,3
RISK	1696	0,4%	0,4%	0,2%	1060	16,7%	1318,6
TUMOR	1673	0,4%	0,4%	0,2%	917	14,4%	1406,0
POSITIVE	1563	0,4%	0,3%	0,2%	905	14,3%	1322,5
PAPILLOMAVIRUS	1524	0,4%	0,3%	0,2%	1367	21,5%	1016,5

Εικόνα 25 Συχνότεροι Όροι που εμφανίζονται στην βιβλιογραφία του όρου "E6 or E7 oncoprotein"

Πίνακας 6

Συχνά Εμφανιζόμενοι Όροι στη Βιβλιογραφία του όρου “E6 or E7 oncoprotein”

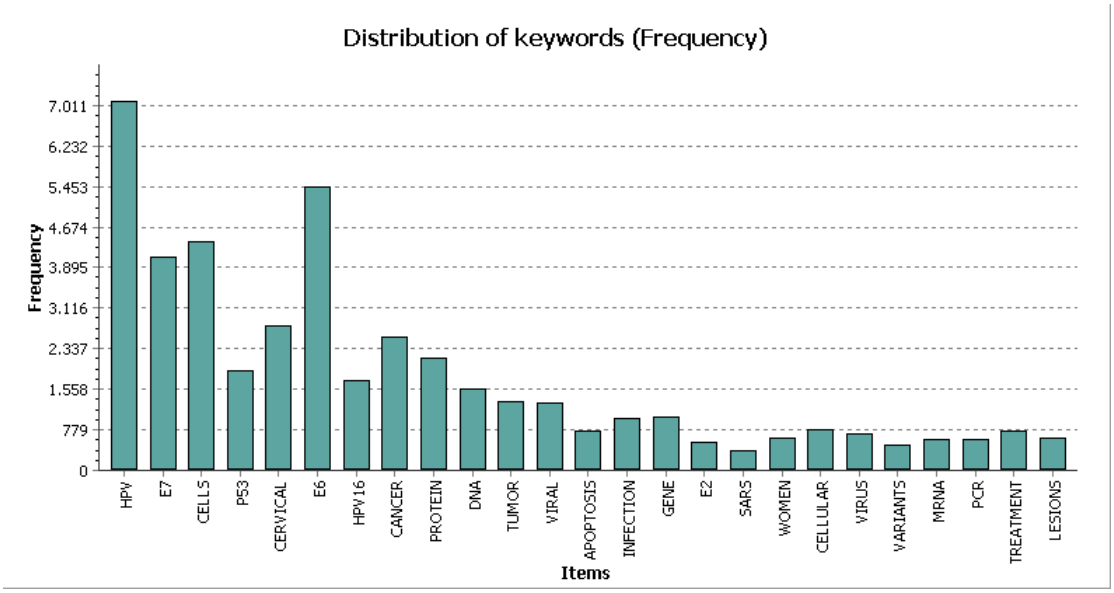
	FREQUENCY	NO. CASES	% CASES	TF • IDF
HPV	9393	2594	40,90%	3652
E6	6762	2832	44,60%	2371,3
CELLS	5547	2086	32,90%	2681,7
E7	5128	2017	31,80%	2554,1
CELL	4300	2061	32,50%	2101,4
CERVICAL	3693	1659	26,10%	2152,8
HUMAN	3556	2402	37,80%	1501,3
EXPRESSION	3373	1675	26,40%	1952,2
CANCER	3279	1685	26,50%	1889,3
PROTEIN	2650	1327	20,90%	1801,7
HPV16	2323	954	15,00%	1912,4
P53	2287	736	11,60%	2140,4
PATIENTS	2199	1208	19,00%	1584,8
DNA	2162	1072	16,90%	1670,3
HIGH	1965	1319	20,80%	1341,2
STUDY	1703	1421	22,40%	1107,3
RISK	1696	1060	16,70%	1318,6
TUMOR	1673	917	14,40%	1406
POSITIVE	1563	905	14,30%	1322,5
PAPILLOMAVIRUS	1524	1367	21,50%	1016,5

Πίνακας 7

Επαναπροσδιορισμένη Λίστα Συχνότερων Όρων της Βιβλιογραφίας του όρου ”e6 or e7 oncoprotein”

	FREQUENCY	NO. CASES	% CASES	TF • IDF
HPV	7081	1633	48,00%	2256,2
E7	4101	1385	40,70%	1600
CELLS	4383	1483	43,60%	1579,9
P53	1916	554	16,30%	1510
CERVICAL	2775	1110	32,60%	1349,4
E6	5447	1970	57,90%	1291,7
HPV16	1727	618	18,20%	1279
CANCER	2562	1211	35,60%	1149
PROTEIN	2158	1005	29,60%	1142,5
DNA	1562	687	20,20%	1085

	FREQUENCY	NO. CASES	% CASES	TF • IDF
TUMOR	1309	670	19,70%	923,5
VIRAL	1285	694	20,40%	887
APOPTOSIS	755	347	10,20%	748,4
INFECTION	1007	623	18,30%	742,3
GENE	1014	633	18,60%	740,4
E2	547	184	5,40%	692,9
SARS	370	63	1,90%	640,9
WOMEN	610	308	9,10%	636,3
CELLULAR	778	531	15,60%	627,5
VIRUS	714	450	13,20%	627,2
VARIANTS	477	165	4,90%	626,8
MRNA	587	291	8,60%	626,7
PCR	599	309	9,10%	623,9
TREATMENT	757	510	15,00%	623,8
LESIONS	630	371	10,90%	606,2



Εικόνα 26 Γραφική Αναπαράσταση Αποτελεσμάτων του Πίνακα 7

Εφαρμόζοντας την ίδια μεθοδολογία για τον όρο **A**: **Loop Electrosurgical Excision**, με τα ίδια φίλτρα αναζήτησης και επεξεργασίας των δεδομένων, παίρνουμε τον **Πίνακα 8** και τη γραφική αναπαράσταση των αποτελεσμάτων στην **Εικόνα 27**.

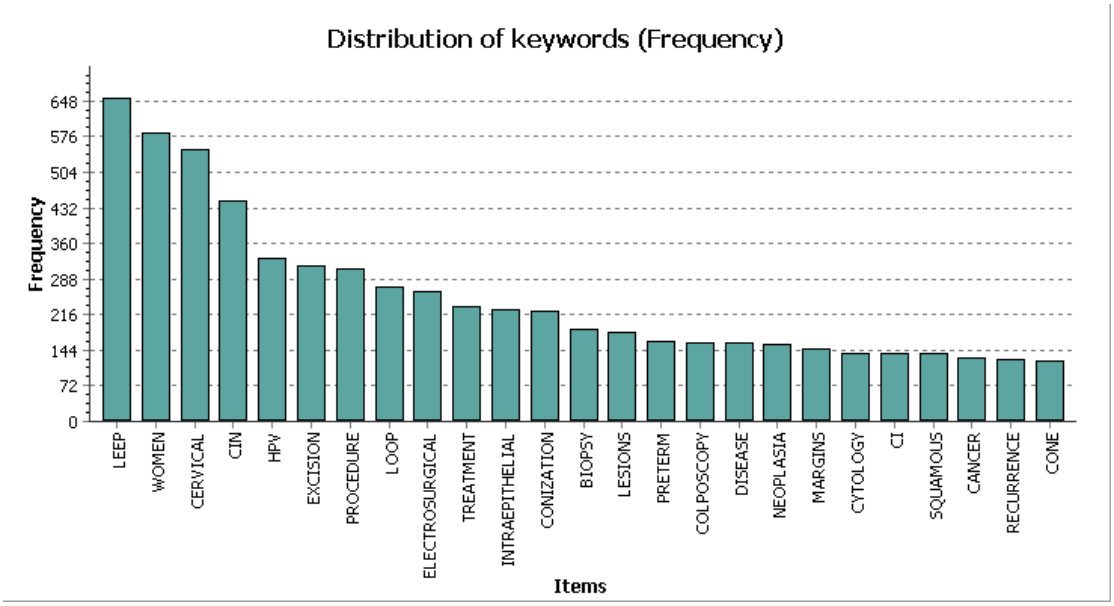
Πίνακας 8

Συχνά Εμφανιζόμενοι Όροι στη Βιβλιογραφία του όρου “Leep Electrosurgical Excision”

	FREQUENCY	NO. CASES	% CASES	TF • IDF
HPV	329	129	15,40%	267,4
CIN	447	217	25,90%	262,3
WOMEN	584	326	38,90%	239,5
LEEP	653	405	48,30%	206,2
CERVICAL	549	371	44,30%	194,3
CI	139	47	5,60%	173,9
CONIZATION	223	148	17,70%	167,9
TREATMENT	231	164	19,60%	163,6
PRETERM	161	84	10,00%	160,8
INTRAEPITHELIAL	225	177	21,10%	151,9
EXCISION	314	276	32,90%	151,5
BIOPSY	186	129	15,40%	151,2
PROCEDURE	309	274	32,70%	150
LESIONS	180	129	15,40%	146,3
MARGINS	148	92	11,00%	142
DISEASE	158	107	12,80%	141,2
LOOP	272	261	31,10%	137,8
ELECTROSURGICAL	263	255	30,40%	135,9
SQUAMOUS	137	86	10,30%	135,5
COLPOSCOPY	159	123	14,70%	132,5
CYTOLOGY	139	97	11,60%	130,2
CANCER	127	84	10,00%	126,9
RECURRENCE	125	86	10,30%	123,6
MARGIN	114	70	8,40%	122,9

Αφού έχουμε πλέον τις λίστες με τους συχνότερους όρους που εμφανίζονται στις δύο βιβλιογραφίες, προχωράμε στην σύγκριση μεταξύ τους. Ενδεικτικά, για τους πρώτους 20 όρους, όπως φαίνεται και από τον **Πίνακα 9** έχουμε 6 κοινούς όρους (**HPV**, **Cervical**, **Cancer**, **Treatment**, **Women, Lesions**) και 1 όρο με παραπλήσια σημασία (**Infection** – **Disease**) οι οποίοι μπορεί να αποτελούν την ενδιάμεση βιβλιογραφία **B** που συνδέει τις βιβλιογραφίες **A** και **C**. Αφαιρώντας τις συνηθισμένες λέξεις (**treatment**, **women**, **infection**) που δεν δίνουν κάποια συγκεκριμένη ιατρική πληροφορία, οι εναπομείναντες όροι (**HPV, Cervical, Cancer, Lesions**) αποτελούν πιθανούς όρους μιας ενδιάμεσης βιβλιογραφίας **B**. Ειδικότερα οι όροι **HPV** και **Cervical**, που καταλαμβάνουν υψηλές θέσεις εμφάνισης συχνότητας και στις δυο βιβλιογραφίες, οδηγούν στο ήδη γνωστό συμπέρασμα, ότι η παρουσία των ογκοπρωτεινών E6, E7 σχετίζεται με τη μέθοδο LEEP, υπό τη γενική μορφή ενός σχεσιακού πλαισίου «αιτία – αντιμετώπιση»: η ανίχνευση των συγκεκριμένων πρωτεϊνών συνδέεται με την εμφάνιση καρκίνου του τραχήλου της μήτρας, που αντιμετωπίζεται με τη μέθοδο LEEP. Συγκρίνοντας τις λίστες με όλους τους συχνά εμφανιζόμενους όρους που εντοπίζονται στις βιβλιογραφίες,

μπορούμε να πάρουμε περισσότερα κοινά αποτελέσματα και να εξετάσουμε περισσότερες πιθανές σχέσεις μεταξύ των δυο βιβλιογραφιών.



Εικόνα 27 Γραφική Αναπαράσταση Αποτελεσμάτων του Πίνακα 8

Πίνακας 9

Σύγκριση 25 πρώτων συχνότερων όρων στις βιβλιογραφίες A, C

	Βιβλιογραφία C	Βιβλιογραφία A
1	HPV	HPV
2	E7	CIN
3	CELLS	WOMEN
4	P53	LEEP
5	CERVICAL	CERVICAL
6	E6	CI
7	HPV16	CONIZATION
8	CANCER	TREATMENT
9	PROTEIN	PRETERM
10	DNA	INTRAEPITHELIAL
11	TUMOR	EXCISION
12	VIRAL	BIOPSY
13	APOPTOSIS	PROCEDURE
14	INFECTION	LESIONS
15	GENE	MARGINS
16	E2	DISEASE
17	SARS	LOOP
18	WOMEN	ELECTROSURGICAL
19	CELLULAR	SQUAMOUS

	Βιβλιογραφία C	Βιβλιογραφία A
20	VIRUS	COLPOSCOPY
21	VARIANTS	CYTOLOGY
22	MRNA	CANCER
23	PCR	RECURRENCE
24	TREATMENT	MARGIN
24	LESIONS	CONE

4.3.3 Επέκταση της Μεθοδολογίας

Το μοντέλο υποθέσεων του Swanson αναφέρεται στη σύγκριση δυο ασύνδετων φαινομενικά βιβλιογραφιών προς εντοπισμό μιας πιθανής κοινής βιβλιογραφίας που να τις συνδέει. Επεκτείνοντας το σύνολο των αρχικών βιβλιογραφιών σε περισσότερες από δυο, μπορούμε να εξάγουμε μια πιο ειδική σχέση που να συνδέει τρείς ή και περισσότερους όρους, πάνω σε ένα συγκεκριμένο ιατρικό θέμα.

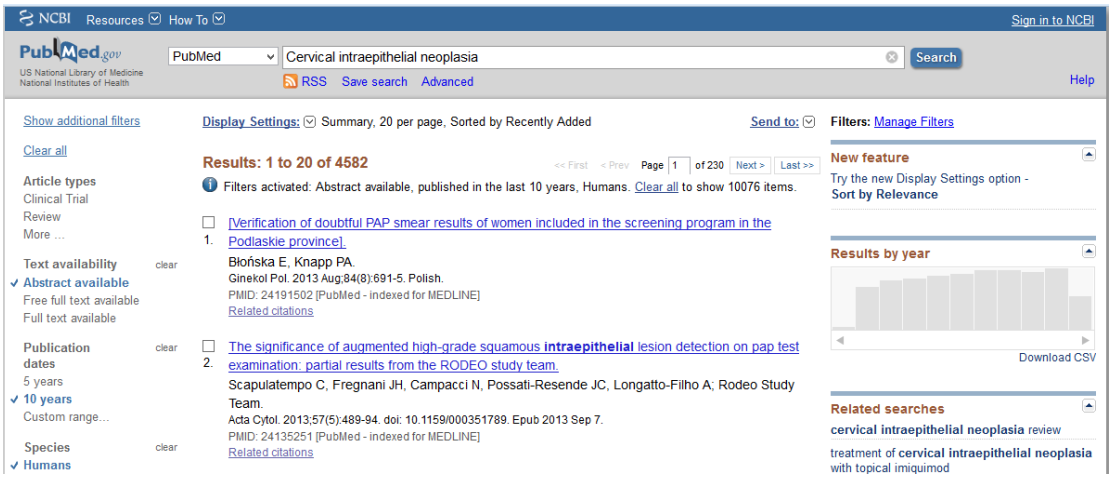
Στα πλαίσια της παραπάνω λογικής και εκμεταλλευόμενοι τα αποτελέσματα της προηγούμενης παραγράφου, αναζητούμε μια επιπλέον βιβλιογραφία για τον όρο

- **D** : **cervical intraepithelial neoplasia**,

ο οποίος αναφέρεται σε δυνητικά προκαρκινικούς μετασχηματισμούς και την ανώμαλη ανάπτυξη (δυσπλασία) των πλακωδών κυττάρων στην επιφάνεια του τραχήλου της μήτρας.

Εφαρμόζοντας τα ίδια φίλτρα αναζήτησης που χρησιμοποιήσαμε στις προηγούμενες δυο βιβλιογραφίες (**Text Availability**: Abstract available, **Publication dates**: 10 years, **Species**: Human), η αναζήτηση επιστρέφει 4582 άρθρα που σχετίζονται με τον συγκεκριμένο όρο (**Εικόνα 29**). Κάνοντας χρήση των λογισμικών **QDA Miner**© και **WordStat**©, παίρνουμε τον **Πίνακα 10** και τη γραφική αναπαράσταση των αποτελεσμάτων στην **Εικόνα 28**.

Εικόνα 28 Γραφική Αναπαράσταση Αποτελεσμάτων του Πίνακα 10



Εικόνα 29 Αποτελέσματα αναζήτησης για τον όρο "Cervical Intraepithelial Neoplasia"

Πίνακας 10

Συχνά Εμφανιζόμενοι Όροι στη Βιβλιογραφία του όρου "Cervical Intraepithelial Neoplasia"

	FREQUENCY	NO. CASES	% CASES	TF • IDF
HPV	17837	5382	41,40%	6823,8
CIN	11244	4348	33,50%	5343,4
WOMEN	10238	5074	39,10%	4178,7
CERVICAL	15115	7341	56,50%	3744,8
CANCER	6069	3299	25,40%	3611,8
LESIONS	4617	2777	21,40%	3093,1
SQUAMOUS	3842	2238	17,20%	2933,9
CYTOLOGY	3824	2339	18,00%	2846,9
SCREENING	3597	2110	16,20%	2838,8
INTRAEPITHELIAL	5411	3964	30,50%	2788,7
CI	2069	673	5,20%	2659,7
CELLS	2769	1605	12,40%	2514,3
NEOPLASIA	3940	3181	24,50%	2407,1
INFECTION	2661	1691	13,00%	2356
DNA	2503	1501	11,60%	2345,6
PAP	2330	1366	10,50%	2278,9
COLPOSCOPY	2341	1619	12,50%	2116,9
TESTING	2107	1346	10,40%	2074,3
HSIL	1741	875	6,70%	2039,6
HR	1569	694	5,30%	1996
TREATMENT	1894	1263	9,70%	1916,9
LSIL	1598	851	6,60%	1891,4
PAPILLOMAVIRUS	2384	2139	16,50%	1867,4

Πίνακας 11

Σύγκριση Συχνότερων Όρων από τρεις βιβλιογραφίες

	Βιβλιογραφία C	Βιβλιογραφία A	Βιβλιογραφία D
1	HPV	HPV	HPV
2	E7	CIN	CIN
3	CELLS	WOMEN	WOMEN
4	P53	LEEP	CERVICAL
5	CERVICAL	CERVICAL	CANCER
6	E6	CI	LESIONS
7	HPV16	CONIZATION	SQUAMOUS
8	CANCER	TREATMENT	CYTOLOGY
9	PROTEIN	PRETERM	SCREENING
10	DNA	INTRAEPITHELIAL	INTRAEPITHELIAL
11	TUMOR	EXCISION	CI
12	VIRAL	BIOPSY	CELLS
13	APOPTOSIS	PROCEDURE	NEOPLASIA
14	INFECTION	LESIONS	INFECTION
15	GENE	MARGINS	DNA
16	E2	DISEASE	PAP
17	SARS	LOOP	COLPOSCOPY
18	WOMEN	ELECTROSURGICAL	TESTING
19	CELLULAR	SQUAMOUS	HSIL
20	VIRUS	COLPOSCOPY	HR
21	VARIANTS	CYTOLOGY	TREATMENT
22	MRNA	CANCER	LSIL
23	PCR	RECURRENCE	PAPILLOMAVIRUS
24	TREATMENT	MARGIN	CARCINOMA
25	LESIONS	CONE	HIV

Από τη σύγκριση των τριών βιβλιογραφιών, παρατηρούμε ότι υπάρχουν 2 ξεχωριστοί κοινοί όροι στη βιβλιογραφία **DC** (σημειωμένοι με κίτρινο), 4 ξεχωριστοί κοινοί όροι (σημειωμένοι με πορτοκαλί) στη βιβλιογραφία **AD,** και 6 κοινοί όροι (σημειωμένοι με πράσινο) στην τομή των βιβλιογραφιών **ACD**. Οι τελευταίοι όροι είναι οι εξής:

"**HPV**, **Cervical**, Cancer, Women, Treatment, Lesions",

όπου διαπιστώνουμε ότι όλοι οι όροι είναι ακριβώς ίδιοι με αυτούς που είχαμε βρει στα αποτελέσματα αναζήτησης πιθανής σύνδεσης των βιβλιογραφιών **A** και **C**, και μάλιστα οι δυο πρώτοι όροι (**HPV**, **Cervical**) κατέχουν εξίσου υψηλή θέση όσον αφορά τη συχνότητα εμφάνισης τους στις τρεις βιβλιογραφίες. Το αποτέλεσμα αυτό υποδηλώνει ότι οι τρεις όροι **A**, **C** και **D** έχουν κοινό υπόβαθρο και συνδέονται λόγω του ιού HPV και της περιοχής του τραχήλου της μήτρας, επαληθεύοντας με αυτόν τον τρόπο την ιατρική υπόθεση που είχαμε διατυπώσει στην αρχή, την ισχύ της οποίας γνωρίζαμε εκ των προτέρων.

4.4 Διερεύνηση Ιατρικής Υπόθεσης: Ιός HPV και Καρκίνος των Πνευμόνων

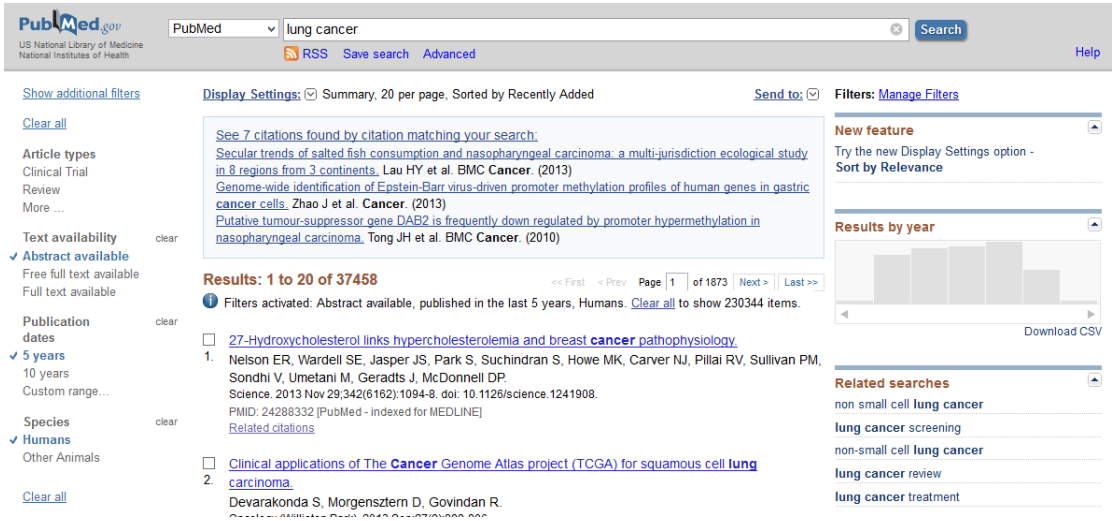
Σύμφωνα με μια πρόσφατη έρευνα[101], ο ιός των ανθρωπίνων θηλωμάτων είναι πιθανό να ευθύνεται για κάποιες μορφές καρκίνου των πνευμόνων. Βάσει αυτής της υπόθεσης, θα εφαρμόσουμε την προηγούμενη μεθοδολογία, ώστε να εξετάσουμε κατά πόσο είναι ή όχι βάσιμη η υπόθεση αυτή.

Θεωρούμε σαν όρους Α και C τους παρακάτω:

- **A**: **Human Papillomavirus**

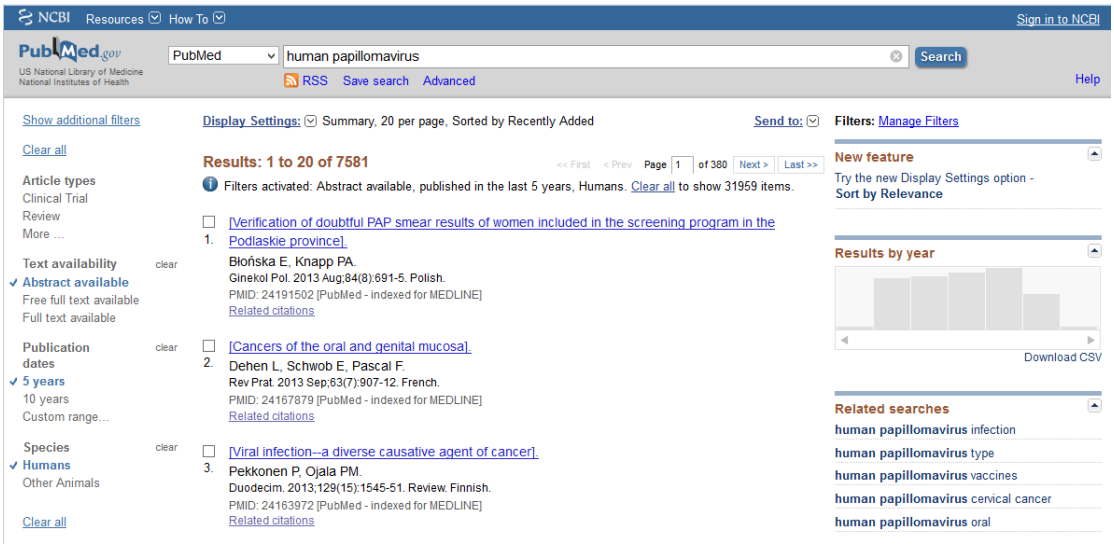
- **C**: **Lung Cancer**

Δεδομένης της ήδη υπάρχουσας γνώσης σχετικά με τους παραπάνω δυο όρους, αναμένεται ένα αρκετά μεγάλο πλήθος από διαθέσιμα άρθρα και δημοσιεύσεις για τις δυο βιβλιογραφίες. Τοποθετώντας τους παραπάνω όρους στη μηχανή αναζήτησης του **PubMed©**, παίρνουμε τα αποτελέσματα που εμφανίζονται στις **Εικόνες 30** και **31** (αφού έχουμε εφαρμόσει, όπως και στις προηγούμενες περιπτώσεις, φίλτρα αναζήτησης: **Text Availability**: Abstract available, **Publication dates**: 5 years, **Species**: Human).



Εικόνα 30 Αποτελέσματα Αναζήτησης για τον όρο "Lung Cancer"

Παρατηρούμε ότι το πλήθος των αποτελεσμάτων για τον όρο **C**: **Lung Cancer** (37458 δημοσιεύσεις και άρθρα) είναι αρκετά μεγάλο για διαχείριση και επεξεργασία πληροφορίας, οπότε αποφασίζουμε να χωρίσουμε τα αποτελέσματα της τελευταίας 5ετίας ανά έτος και να πάρουμε με αυτό τον τρόπο 5 αρχεία .XML, το καθένα για τα έτη από το 2009 έως το 2013.



Εικόνα 31 Αποτελέσματα Αναζήτησης για τον Όρο "Human Papillomavirus"

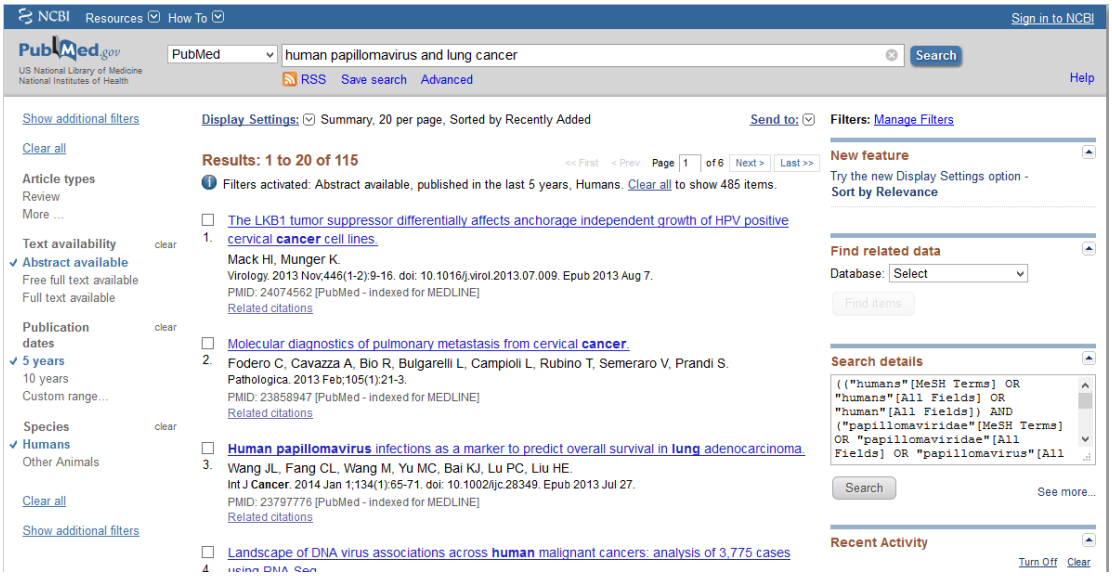
Επιπλέον, για να έχουμε μια πιο αυθεντική και μη βεβιασμένη ιατρική υπόθεση, κάνουμε μια επιπλέον αναζήτηση στο **PubMed©**, ώστε να ελέγξουμε αν υπάρχει κοινή βιβλιογραφία για τους όρους "Human Papillomavirus" και "Lung Cancer", με σκοπό να την αποσύρουμε από την διερεύνηση που θα υλοποιήσουμε. Για το χρονικό διάστημα των 5 χρόνων που εξετάζουμε, η αναζήτηση επιστρέφει 115 άρθρα, όπως φαίνεται στην **Εικόνα 32**, τα οποία και θα αγνοήσουμε, εφαρμόζοντας εκ νέου φίλτρα στις μεμονωμένες αναζητήσεις των όρων "Human Papillomavirus" και "Lung Cancer".

Το **PubMed©**, προσφέρει μέσω της επιλογής **Advanced** (**Εικόνα 33**) την εφαρμογή εξειδικευμένων φίλτρων, οπότε ορίζουμε τις επιλογές αναζήτησης με τέτοιο τρόπο, ώστε να υπολογίσουν τα αποτελέσματα για τους όρους "Human Papillomavirus" και "Lung Cancer", αγνοώντας την μεταξύ τους κοινή βιβλιογραφία. Όπως φαίνεται και στην **Εικόνα 34**, τα φίλτρα που εφαρμόζουμε έχουν τη μορφή:

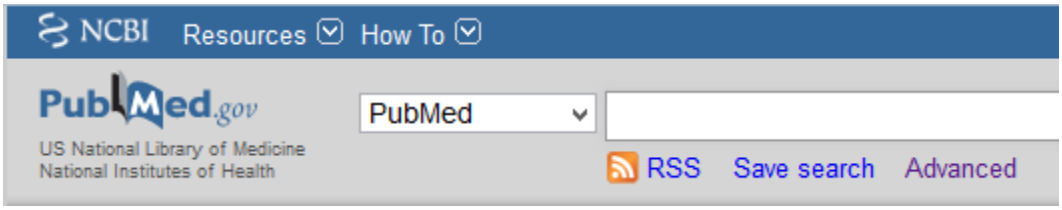
- lung cancer AND hasabstract[text] AND "last 5 years"[PDat] AND Humans[Mesh]
- NOT human papillomavirus and lung cancer AND hasabstract[text] AND "last 5 years"[PDat] AND Humans[Mesh]

και

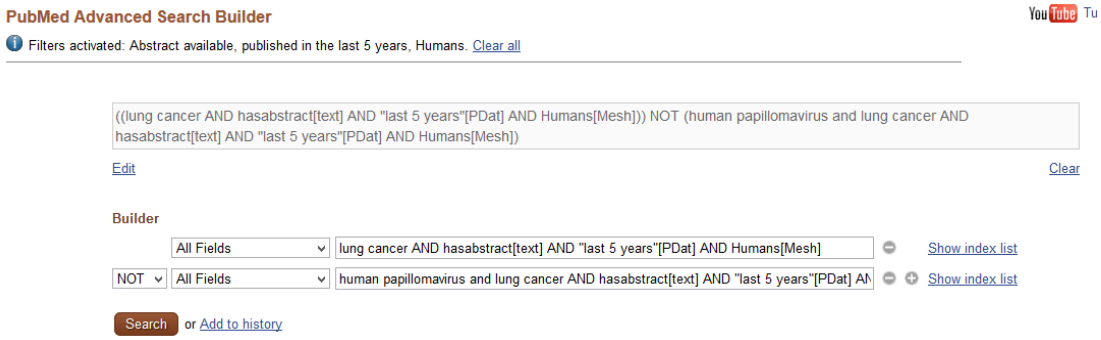
- human papillomavirus AND hasabstract[text] AND "last 5 years"[PDat] AND Humans[Mesh]
- NOT human papillomavirus and lung cancer AND hasabstract[text] AND "last 5 years"[PDat] AND Humans[Mesh]



Εικόνα 32 Αποτελέσματα Αναζήτησης για κοινή βιβλιογραφία των όρων "Human Papillomavirus" και "Lung Cancer"



Εικόνα 33 Επιλογή **Advanced**, στην μηχανή αναζήτησης του PubMed©



Εικόνα 34 Εφαρμογή φίλτρων για την αναζήτηση βιβλιογραφίας, χωρίς καμία κοινή αναφορά των όρων "Human Papillomavirus" και "Lung Cancer"

Μετά την εφαρμογή των παραπάνω φίλτρων, τα αποτελέσματα των αναζητήσεων για βιβλιογραφίες των όρων **A** και **C**, επιστρέφουν 7446 και 37343 αρχεία αντίστοιχα, επαληθεύοντας την αφαίρεση της κοινής βιβλιογραφίας με τα προηγούμενα αποτελέσματα

της αναζήτησης (7581 – 115 = 7446 και 37458 - 115 = 37343). Αφού αποθηκεύσουμε τα 5 αρχεία .XML, που αναφέρονται στον όρο “Lung Cancer” και το αρχείο .XML για τον όρο “Human Papillomavirus”, προχωράμε στην επεξεργασία των δεδομένων με τη βοήθεια των προγραμμάτων **QDA Miner©** και **WordStat©**.

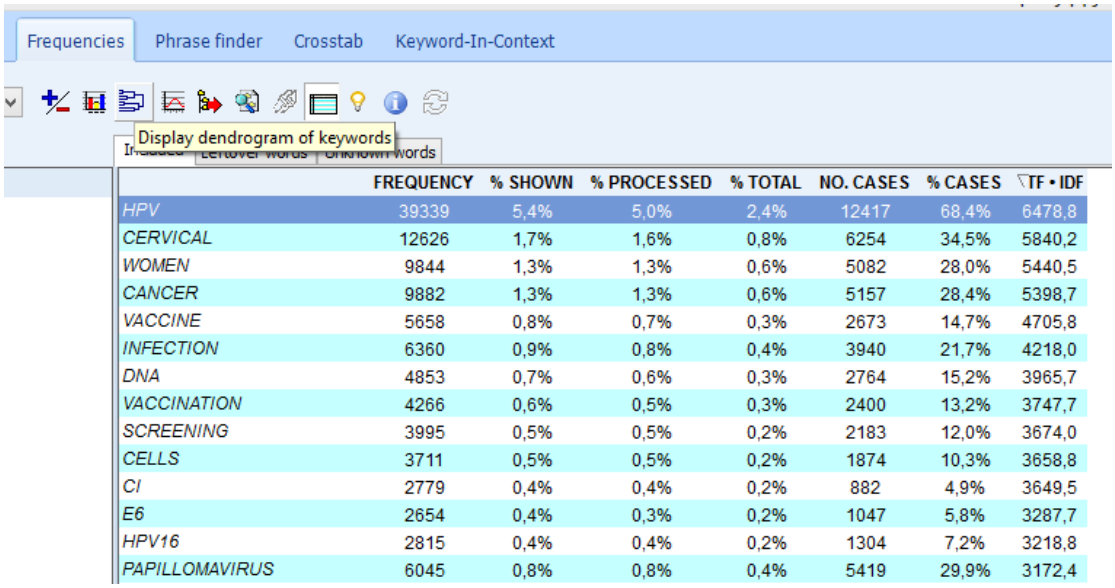
Στο σημείο αυτό, θα διαφοροποιήσουμε τη μεθοδολογία που παρουσιάστηκε στην παράγραφο 4.3.1, διότι στην διερεύνηση που θα υλοποιήσουμε, δεν έχουμε εκ των προτέρων γνώση για την ύπαρξη κοινής βιβλιογραφίας ανάμεσα στις δυο βιβλιογραφίες των όρων **A** και **C**, όπως είχαμε με την εφαρμογή των προηγούμενων παραδειγμάτων και επιπλέον αποκλείσαμε όποια κοινή βιβλιογραφία υπήρχε, ώστε να έχουμε όσο το δυνατόν περισσότερο, φαινομενικά ασύνδετες βιβλιογραφίες. Κατά συνέπεια, αυτό που μας ενδιαφέρει είναι εάν υπάρχουν όροι της βιβλιογραφίας **A** που να ανιχνεύονται στην βιβλιογραφία **C** ώστε να δώσουν πιθανούς όρους **B**, και όχι η συχνότητα εμφάνισης των όρων στις δυο βιβλιογραφίες, που χρησιμοποιήθηκε στις προηγούμενες εφαρμογές σαν μέτρο σύγκρισης για να αποδείξουμε την ισχυρή σύνδεση των βιβλιογραφιών που εξετάσαμε.

4.4.1 Παρουσίαση Μεθοδολογίας και Υλοποίηση

Δεδομένου ότι ψάχνουμε για πιθανές σχέσεις του ιού HPV με την εμφάνιση κάποιων μορφών του καρκίνου των πνευμόνων και έχοντας απορρίψει οποιαδήποτε άμεση σχέση υπήρχε στις δυο βιβλιογραφίες, θα αναζητήσουμε αρχικά λέξεις-κλειδιά που συνδέονται άρρηκτα με τον ιό HPV και στη συνέχεια θα εξετάσουμε αν υπάρχουν πιθανές σχέσεις αυτών των όρων με τη βιβλιογραφία του καρκίνου των πνευμόνων.

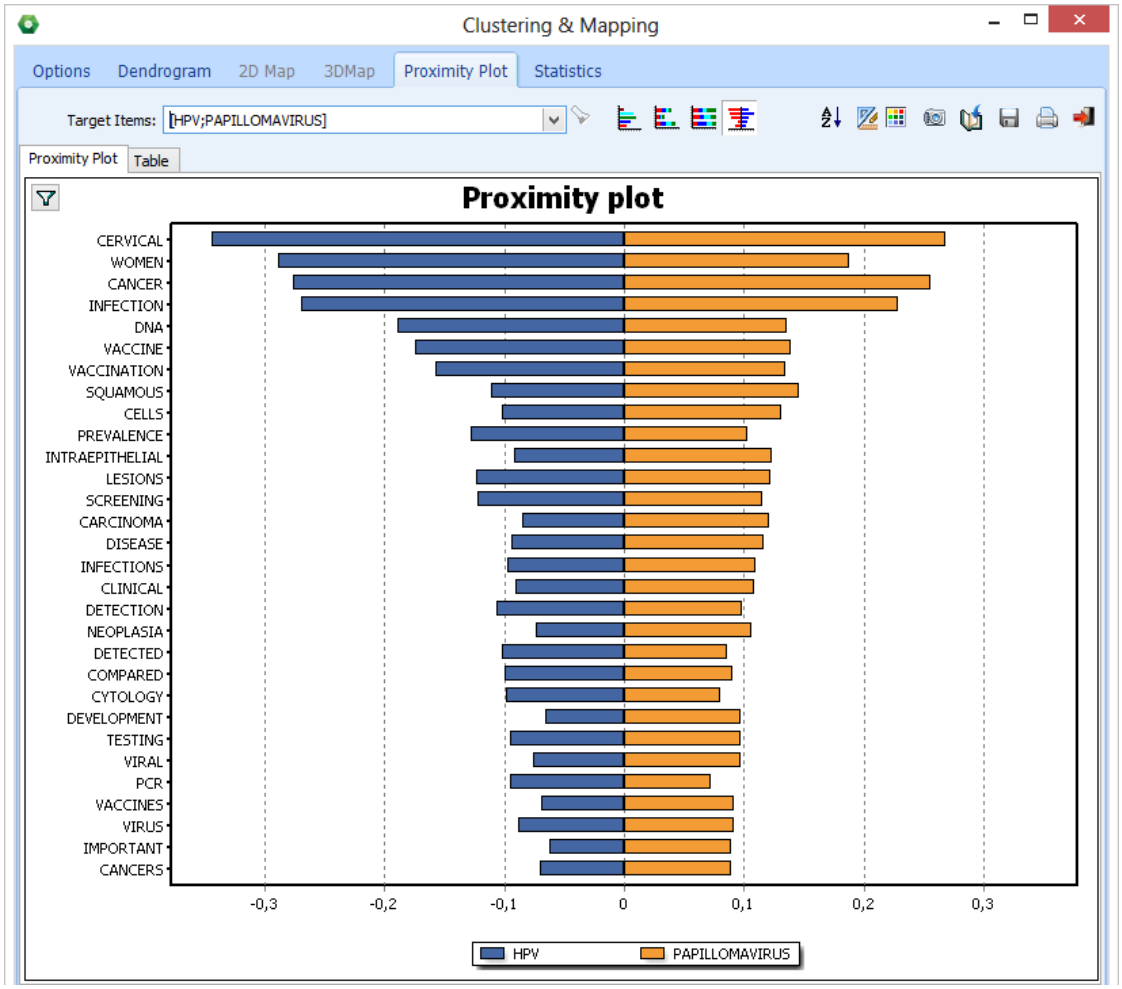
Ένας τρόπος για να υλοποιηθεί η παραπάνω πρόταση, είναι να επεξεργαστούμε τα αποτελέσματα της βιβλιογραφίας του όρου HPV και να εντοπίσουμε για παράδειγμα, τους 5 έως 10 πιο συχνά εμφανιζόμενους όρους που συνδέονται με τον HPV. Αφού εισάγουμε τα δεδομένα στο **QDA Miner©**, προχωράμε στην επεξεργασία τους με το **WordStat©**. Για να βρούμε τους όρους που σχετίζονται περισσότερο με τον ιό HPV, χρησιμοποιούμε την επιλογή Display Dendrogram of keywords (**Εικόνα 35**) και επιλέγουμε τη σχεδίαση ενός διαγράμματος γειτνίασης (**Proximity Plot**) για τους όρους “**HPV**“ και “**Papillomavirus**” (που ουσιαστικά αναφέρονται στον ίδιο όρο, τον ιό των ανθρωπίνων θηλωμάτων), για να διαπιστώσουμε ποιες λέξεις-κλειδιά σχετίζονται περισσότερο με αυτούς τους συγκεκριμένους όρους. Από το διάγραμμα γειτνίασης (**Εικόνα 36**) και έχοντας υπόψιν να μην συμπεριλάβουμε λέξεις που δεν προσφέρουν ιατρική πληροφορία (women, cells, prevalence, ..) ή αναμενόμενα γνωστή ιατρική πληροφορία (cancer, carcinoma, dna, ..), καταλήγουμε στην παρακάτω λίστα με 6 λέξεις-κλειδιά:

$$\lambda = \{\textbf{cervical}, \textbf{squamous}, \textbf{intraepithelial}, \textbf{lesions}, \textbf{cytology}, \textbf{pcr}\}$$



	FREQUENCY	% SHOWN	% PROCESSED	% TOTAL	NO. CASES	% CASES	TF • IDF
HPV	39339	5,4%	5,0%	2,4%	12417	68,4%	6478,8
CERVICAL	12626	1,7%	1,6%	0,8%	6254	34,5%	5840,2
WOMEN	9844	1,3%	1,3%	0,6%	5082	28,0%	5440,5
CANCER	9882	1,3%	1,3%	0,6%	5157	28,4%	5398,7
VACCINE	5658	0,8%	0,7%	0,3%	2673	14,7%	4705,8
INFECTION	6360	0,9%	0,8%	0,4%	3940	21,7%	4218,0
DNA	4853	0,7%	0,6%	0,3%	2764	15,2%	3965,7
VACCINATION	4266	0,6%	0,5%	0,3%	2400	13,2%	3747,7
SCREENING	3995	0,5%	0,5%	0,2%	2183	12,0%	3674,0
CELLS	3711	0,5%	0,5%	0,2%	1874	10,3%	3658,8
CI	2779	0,4%	0,4%	0,2%	882	4,9%	3649,5
E6	2654	0,4%	0,3%	0,2%	1047	5,8%	3287,7
HPV16	2815	0,4%	0,4%	0,2%	1304	7,2%	3218,8
PAPILLOMAVIRUS	6045	0,8%	0,8%	0,4%	5419	29,9%	3172,4

Εικόνα 35 Αποτελέσματα επεξεργασίας δεδομένων για τον όρο HPV και σχεδίαση δενδρογράμματος



Εικόνα 36 Διάγραμμα Γειτνίασης των όρων HPV και Papillomavirus

Στη συνέχεια εξετάζουμε εάν οι όροι της λίστας **λ** συναντώνται στη βιβλιογραφία **C**, ώστε να δώσουν πιθανούς όρους **B**. Λόγω της ύπαρξης 5 αρχείων .XML που περιέχουν τη συγκεκριμένη βιβλιογραφία, θα χρειαστούν 5 διαφορετικοί έλεγχοι, ώστε να καλυφθεί ολόκληρη η βιβλιογραφία. Στη συγκεκριμένη επεξεργασία για τη βιβλιογραφία του όρου "Lung Cancer" δεν θα θέσουμε κάποιο όριο στους εμφανιζόμενους όρους, ούτε θα τους κατατάξουμε με κάποιο φίλτρο συχνότητας εμφάνισης, διότι μας ενδιαφέρει μόνο η πιθανή ταυτοποίηση των όρων της λίστας **λ** με τους όρους της βιβλιογραφίας **C**.

Αφού πάρουμε τα αποτελέσματα από την επεξεργασία των 5 αρχείων .XML στο **WordStat©**, τα εξάγουμε υπό τη μορφή 5 πινάκων δεδομένων στο **Microsoft® Office Excel®** 2010, όπου θα υλοποιήσουμε την αναζήτηση και ταυτοποίηση των όρων της λίστας **λ** με τα δεδομένα των πινάκων. Για την ταυτοποίηση, θα χρησιμοποιήσουμε την συνάρτηση που υπάρχει στο **Excel®**:

Match (Lookup_value; Lookup_array; Match_type)

όπου

- **Lookup_value**: είναι η τιμή που θέλουμε να αντιστοιχίσουμε στο Lookup_array
- **Lookup_array**: είναι το εύρος των δεδομένων στο οποίο θα γίνει η αναζήτηση
- **Match_type**: με ποιο τρόπο θα γίνει η αντιστοίχιση της τιμής του πεδίου Lookup_value στο Lookup_array

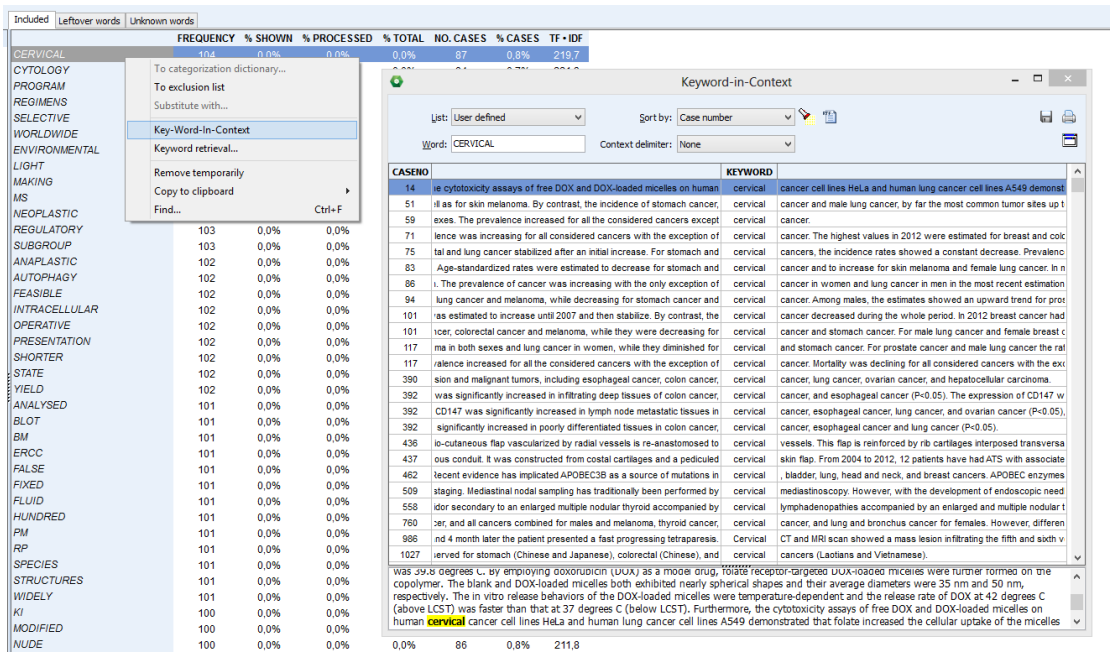
Στο αρχείο επεξεργασίας δεδομένων του **Excel®**, θα τοποθετήσουμε στην πρώτη στήλη τους όρους της λίστας **λ**, ενώ στις επόμενες 5 στήλες, θα τοποθετήσουμε τους όρους που προέκυψαν από την ανάλυση κειμένου των 5 αρχείων .XML της βιβλιογραφίας **C**. Στη συνέχεια, θα εφαρμόσουμε τη συνάρτηση **match**, όπου στο πεδίο **lookup_value** θα χρησιμοποιήσουμε τους όρους της λίστας **λ**, στο πεδίο **lookup_array**, θα επιλέξουμε κάθε φορά μια από τις επόμενες 5 στήλες, που περιέχουν τους όρους των 5 ετών από τη βιβλιογραφία **C** και στο πεδίο **match_type**, θα επιλέξουμε την τιμή 0, ώστε να μας επιστρέψει τη θέση που κατέχει ο όρος από το πεδίο **lookup_value** στο εύρος τιμών του **lookup_array**. Στην **Εικόνα 37** φαίνονται τα αποτελέσματα από την παραπάνω εφαρμογή ανάλυσης των δεδομένων.

	A	B	C	D	E	F	G	H	I	J	K
1	HPV	2013	2012	2011	2010	2009	match_2013	match_2012	match_2011	match_2010	match_2009
2	cervical	CELLS	CANCER	CANCER	CANCER	CANCER	1011	811	817	795	637
3	squamous	CANCER	CELLS	CELLS	CELLS	CELLS	132	103	105	149	127
4	intraepithelial	TUMOR	LUNG	LUNG	TUMOR	LUNG	5944	#N/A	7721	#N/A	#N/A
5	lesions	LUNG	TUMOR	TUMOR	LUNG	TUMOR	57	56	66	52	39
6	cytology	NSCLC	NSCLC	TREATMENT	TREATMENT	TREATMENT	998	920	992	818	1050
7	pcr	EGFR	TREATMENT	NSCLC	NSCLC	NSCLC	268	203	256	271	265
8		TREATMENT	EGFR	EGFR	DISEASE	DISEASE					
9		CI	CI	DISEASE	EGFR	EGFR					
10		DISEASE	DISEASE	CI	CHEMOTHERAPY	TUMORS					
11		METASTASIS	CT	TUMORS	CT	CT					
12		CLINICAL	METASTASIS	CHEMOTHERAPY	TUMORS	CHEMOTHERAPY					
13		CT	CHEMOTHERAPY	CT	CLINICAL	PULMONARY					
14		TUMORS	CLINICAL	CLINICAL	PULMONARY	CLINICAL					
15		CHEMOTHERAPY	TUMORS	PULMONARY	GROWTH	CI					
16		PULMONARY	PULMONARY	METASTASIS	METASTASIS	CARCINOMA					
17		THERAPY	THERAPY	CARCINOMA	CI	METASTASIS					
18		GROWTH	CARCINOMA	THERAPY	THERAPY	GROWTH					
19		BREAST	GROWTH	GROWTH	CARCINOMA	METASTASES					
20		CARCINOMA	COMPARED	METASTASES	GENE	BREAST					
21		COMPARED	GENE	COMPARED	METASTASES	THERAPY					

Εικόνα 37 Αποτελέσματα Ανάλυσης Δεδομένων για τις βιβλιογραφίες **A**, **C**

Σύμφωνα με τα παραπάνω αποτελέσματα οι όροι **cervical**, **squamous**, **lesions**, **cytology**, **pcr** συναντώνται και στα 5 έτη της βιβλιογραφίας του όρου "Lung Cancer", ενώ ο όρος **intraepithelial**, συναντάται μόνο τα έτη 2013 και 2011. Τα αποτελέσματα αυτά δείχνουν ότι οι παραπάνω όροι αποτελούν πιθανούς όρους βιβλιογραφίας **B**, που συνδέουν τις δυο ασύνδετες βιβλιογραφίες **A**, **C**, επομένως υπάρχει πιθανή σύνδεση του ιού των ανθρωπίνων θηλωμάτων με τον καρκίνο των πνευμόνων.

Για τον εντοπισμό των κειμένων στα οποία υπάρχει σύνδεση των όρων των δυο βιβλιογραφιών, χρησιμοποιούμε το λογισμικό **WordStat©**. Έστω ότι η αναζήτηση γίνεται στην βιβλιογραφία **C** του έτους 2013, για αναφορές του όρου **cervical**. Αρχικά πατώντας Ctrl+F, εντοπίζουμε την θέση του όρου **cervical** και στη συνέχεια με δεξί κλικ πάνω στον όρο, επιλέγουμε το **Key_Word_In_Context**, ώστε να μας εμφανίσει σε ποιες περιπτώσεις – περιλήψεις άρθρων, εμφανίζεται ο όρος **cervical** μέσα στη βιβλιογραφία του καρκίνου των πνευμόνων (**Εικόνα 38**).



Εικόνα 38 Εντοπισμός όρου "Cervical" στη Βιβλιογραφία του όρου "Lung Cancer"

Από το ίδιο παράθυρο, μπορούμε να αποθηκεύσουμε το αρχείο με τα επιλεγμένα κείμενα περιλήψεων, σε μορφή .txt και να έχουμε με αυτόν τον τρόπο συγκεντρωμένες όλες τις περιλήψεις (**Εικόνα 39**). Σε περίπτωση αντίστροφης αναζήτησης, μπορούμε να εισάγουμε το κείμενο της περίληψης στη μηχανή αναζήτησης του **PubMed©** και να μας εμφανίσει το αντίστοιχο άρθρο ή δημοσίευση, όπως εφαρμόσαμε και για το πρώτο αποτέλεσμα της παραπάνω αναζήτησης (**Εικόνα 40**).

File Edit Format View Help

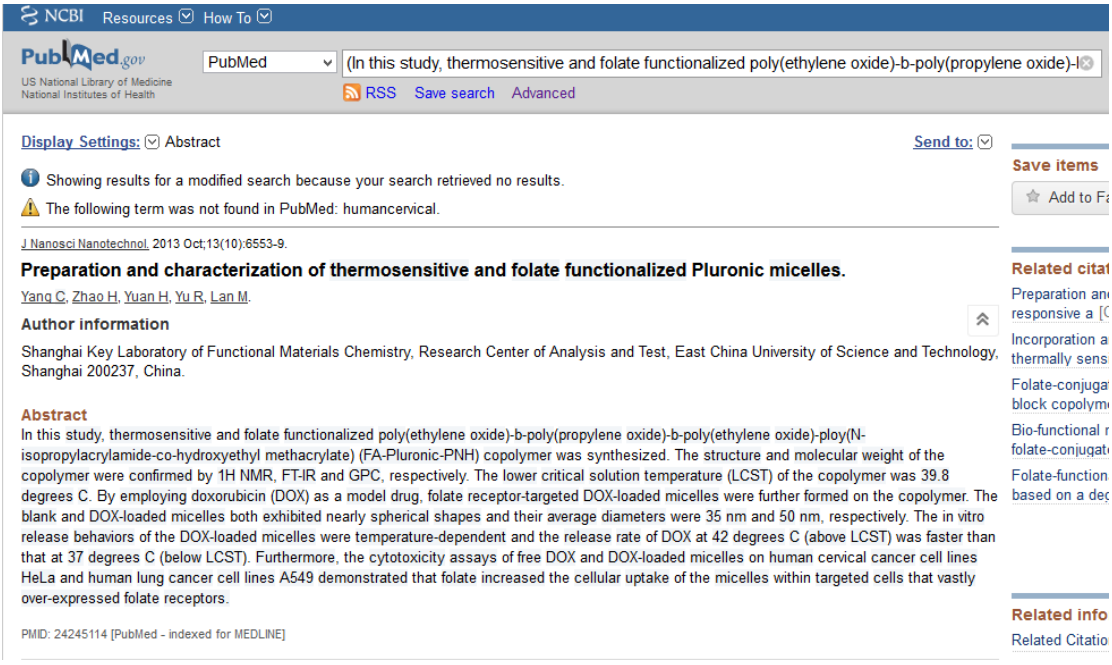
CASENO KEYWORD

14 In this study, thermosensitive and folate functionalized poly(ethylene oxide)-b-poly (propylene oxide)-b-poly(ethylene oxide)-ploy(N-isopropylacrylamide-co-hydroxyethyl methacrylate) (FA-Pluronic-PNH) copolymer was synthesized. The structure and molecular weight of the copolymer were confirmed by 1H NMR, FT-IR and GPC, respectively. The lower critical solution temperature (LCST) of the copolymer was 39.8 degrees C. By employing doxorubicin (DOX) as a model drug, folate receptor-targeted DOX-loaded micelles were further formed on the copolymer. The blank and DOX-loaded micelles both exhibited nearly spherical shapes and their average diameters were 35 nm and 50 nm, respectively. The in vitro release behaviors of the DOX-loaded micelles were temperature-dependent and the release rate of DOX at 42 degrees C (above LCST) was faster than that at 37 degrees C (below LCST). Furthermore, the cytotoxicity assays of free DOX and DOX-loaded micelles on humancervical cancer cell lines HeLa and human lung cancer cell lines A549 demonstrated that folate increased the cellular uptake of the micelles within targeted cells that vastly over-expressed folate receptors.

51 Our findings indicate that breast, colon-rectum and prostate will be the cancer sites with the highest incidence rates in the forthcoming years. The incidence rates still tend to increase for breast, male colorectal cancer and female lung cancer as well as for skin melanoma. By contrast, the incidence of stomach cancer,cervical cancer and male lung cancer, by far the most common tumor sites up to the early 1990s, will continue to decrease. The mortality estimates showed a decreasing trend for all considered cancers with the only exception of lung cancer in women.

59 In 2012 the most common cancers were breast cancer in women, colorectal cancer in both sexes, and prostate cancer in men, with about 4,000, 3,500 and 3,000 estimated new cases, respectively. The highest crude mortality rates were estimated for lung cancer in men (63.6 per 100,000) and breast cancer in women (30.8 per 100,000) and the lowest for skin melanoma (both sexes) and cancer of the cervix uteri. For colorectal, lung and stomach cancer and skin melanoma, all the indicators were higher in men than women. The prevalence figures in women were more than 9 times the incidence figures for breast cancer and more than 10 times the incidence figures for skin melanoma. The prevalence was twice the incidence for lung cancer in both sexes. The prevalence increased for all the considered cancers exceptcervical cancer.

Εικόνα 39 Αρχείο .txt με τις συγκεντρωμένες περιλήψεις για τον όρο “Cervical”



Εικόνα 40 Αντίστροφη αναζήτηση, από το AbstractText στη Δημοσίευση

Κεφάλαιο 5: Συμπεράσματα και Προτάσεις

5.1 Συμπεράσματα

Η παρούσα διπλωματική εργασία, είχε ως στόχο την διερεύνηση μιας συγκεκριμένης ιατρικής υπόθεσης, που αφορά την ύπαρξη ή έλλειψη σχέσης ανάμεσα στον ιό των ανθρωπίνων θηλωμάτων και τον καρκίνο των πνευμόνων.

Χρησιμοποιώντας τεχνικές εξόρυξης γνώσης από κείμενα, απομονώσαμε λέξεις-κλειδιά από τη βιβλιογραφία του ιού HPV για τα τελευταία 5 χρόνια, από το 2009 έως το 2013 και εν συνεχεία, εφαρμόσαμε τις βασικές αρχές του μοντέλου υποθέσεων του Swanson, ώστε να αποτελέσουν αυτές οι λέξεις-κλειδιά τους όρους μιας πιθανής ενδιάμεσης βιβλιογραφίας ανάμεσα στον ιό των ανθρωπίνων θηλωμάτων και τον καρκίνο των πνευμόνων. Από αυτές τις λέξεις-κλειδιά, επιλέξαμε τους όρους που εμφανίζονται συχνότερα στη συγκεκριμένη βιβλιογραφία (και κατά συνέπεια συνδέονται άμεσα με τον ιό HPV) και ανακαλύψαμε ότι οι όροι αυτοί αναφέρονται και στη βιβλιογραφία του καρκίνου των πνευμόνων, από το 2009 έως το 2013.

Σύμφωνα με τα αποτελέσματα που προέκυψαν από την διερεύνηση αυτής της ιατρικής υπόθεσης, οδηγούμαστε στο συμπέρασμα ότι η ύπαρξη σχέσης ανάμεσα στους δυο αυτούς όρους είναι αρκετά πιθανή. Επιπλέον, η παρουσία των 115 άρθρων/δημοσιεύσεων στο **PubMed©** την τελευταία 5ετία, που εν γνώσει μας αποκλείσαμε από την διερεύνηση μας για να μη θεωρηθεί προκατειλημμένη ή ότι έχει υποστεί αλλοίωση λόγω πρωθύστερης γνώσης, επαληθεύει το συμπέρασμα στο οποίο καταλήγουμε και δηλώνει την ύπαρξη μιας κατευθυντήριας τάσης για περαιτέρω έρευνα και μελέτη πάνω στη σχέση αυτών των δύο όρων.

Το αποτέλεσμα αυτό ενισχύεται και από την πρόσφατη γνώση ότι ο ιός HPV ευθύνεται για κάποιους τύπους καρκίνου της στοματοφαρυγγικής κοιλότητας[102], γεγονός που δηλώνει ότι ο ιός των ανθρωπίνων θηλωμάτων ίσως ευθύνεται και για άλλες μορφές καρκίνου και ότι δεν δρα μεμονωμένα στην περιοχή του επιθηλίου του τραχήλου της μήτρας. Δεδομένου ότι μέσω της στοματοφαρυγγικής κοιλότητας υπάρχει σύνδεση με το αναπνευστικό σύστημα, αποκτά περαιτέρω βάση η συσχέτιση που εκφράστηκε στην παρούσα εργασία, για πιθανή σύνδεση του ιού HPV με κάποιους τύπους καρκίνου των πνευμόνων.

5.2 Περιορισμοί της παρούσας εργασίας

Όπως αναφέραμε στην προηγούμενη παράγραφο, για την διερεύνηση της ιατρικής υπόθεσης έγινε χρήση του μοντέλου υποθέσεων του Swanson το οποίο, δοθείσης μια συγκεκριμένης ερώτησης έρευνας στη βιοιατρική, θέτει ως στόχο να εντοπιστούν δύο συμπληρωματικές αλλά ασύνδετες βιβλιογραφίες (**A** και **C**) και επιδιώκει να συνδέσει τις υπάρχουσες γνώσεις από εμπειρικά αποτελέσματα, φέρνοντας στο φως σχέσεις που εμπλέκονται και "αμελούνται", μέσω κοινών για τις δυο βιβλιογραφίες όρων (**B**). Το μοντέλο αυτό, παρόλο που εξακολουθεί να είναι ένα από τα βασικά μοντέλα για την διερεύνηση και μελέτη ιατρικών υποθέσεων, ίσως θεωρηθεί πρωτόλειο και ανεπαρκές για τη διεξαγωγή και εξακρίβωση μιας ιατρικής υπόθεσης με ανάλογη ισχύ. Σαν μειονέκτημα μπορεί να

θεωρηθεί και η χρήση ενός γενικού όρου όπως ο “Lung Cancer”. Πιθανή συγκεκριμενοποίηση και στόχευση σε ειδικότερη ορολογία είναι πιθανό να οδηγήσει σε καλύτερα αποτελέσματα και σε μεγαλύτερη ταυτοποίηση όρων ανάμεσα στις βιβλιογραφίες.

Ακολούθως, η μεθοδολογία που εφαρμόσαμε καθώς και το πλαίσιο αναφοράς που θέσαμε για την διερεύνηση έχει αρκετούς περιορισμούς. Η αναζήτηση σχέσεων ανάμεσα σε όρους που εμφανίζονται συχνότερα σε μια βιβλιογραφία και ο έλεγχος για ύπαρξη τους σε μια δεύτερη βιβλιογραφία, ίσως αποκλείει πιθανές σχέσεις ανάμεσα σε λιγότερο συχνά εμφανιζόμενους όρους, οι οποίοι να οδηγούν σε νέες ανακαλύψεις και θεωρήσεις για την ιατρική υπόθεση. Επιπλέον, η έρευνα περιορίστηκε μόνο σε μια γνωσιακή βάση βιοιατρικών δεδομένων (**PubMed©**) και για ένα χρονικό διάστημα 5 ετών, το οποίο ίσως δεν προσφέρει ολοκληρωμένη κάλυψη της υπάρχουσας βιβλιογραφίας, τόσο για τον ιό HPV όσο και για τον καρκίνο των πνευμόνων. Ο συγκεκριμένος ορισμός του χρονικού διαστήματος της έρευνας περιορίστηκε και από τις υπολογιστικές δυνατότητες των υφιστάμενων εργαλείων ανάλυσης και εξαγωγής πληροφοριών από κείμενα.

Τέλος, η σύνδεση και η εξαγωγή των πιθανών σχέσεων ανάμεσα στις δυο βιβλιογραφίες που μελετήθηκαν στην παρούσα διπλωματική εργασία, πρέπει να εξεταστούν από ιατρικούς ειδήμονες και να εξακριβωθεί η ισχύ τους, μέσα από πειραματικές μελέτες και περαιτέρω έρευνες.

5.3 Προτάσεις για Περαιτέρω Έρευνα

Για την περαιτέρω μελέτη του αντικειμένου της παρούσας Διπλωματικής Εργασίας, ενδιαφέρον θα παρουσίαζαν οι παρακάτω προτάσεις:

I. Η αλλαγή του διαστήματος μελέτης των βιβλιογραφιών των δυο όρων, είναι πιθανό να οδηγήσει στην εύρεση περισσοτέρων κοινών όρων και στην ανακάλυψη νέων σχέσεων ανάμεσα στους δυο όρους.

II. Η άρση περιορισμών στην εύρεση των κοινών όρων στις βιβλιογραφίες μπορεί να οδηγήσει σε ανακάλυψη νέων σχέσεων ανάμεσα σε μη συχνά εμφανιζόμενους όρους.

III. Η χρήση περισσότερων από μία αποθηκών βιοιατρικής βιβλιογραφίας που να διαθέτουν βιοιατρικές έννοιες διαφόρων σημασιολογικών τύπων, μπορούν να επιτρέψουν την ανακάλυψη υποθέσεων μεγάλης κλίμακας, διασταυρωμένων με περισσότερες αποθήκες δεδομένων, ξεπερνώντας τις ανακαλύψεις που βασίζονται στην ανάκτηση πληροφοριών με τα υπάρχοντα εργαλεία ανάκτησης γνώσης από κείμενα.

IV. Η εφαρμογή ενός διαφορετικού μοντέλου υποθέσεων για να εξακριβώσει ή να απορρίψει τα αποτελέσματα που εξήχθησαν με το μοντέλο υποθέσεων του Swanson.

V. Για την διατύπωση της υπόθεσης, τα αποτελέσματα από διαφορετικούς πόρους (για παράδειγμα, βιβλιογραφία και βάσεις δεδομένων) μπορούν να συγκριθούν μεταξύ τους, χρησιμοποιώντας κριτήρια σημασιολογικής ομοιότητας.

VI. Πιο στοχευμένη έρευνα και επιλογή ειδικότερης ορολογίας, είναι πιθανό να οδηγήσει σε καλύτερα αποτελέσματα και σε μεγαλύτερη ταυτοποίηση όρων ανάμεσα στις βιβλιογραφίες.

Παράρτημα

Το λογισμικό **QDA Miner©** παρέχεται από την εταιρεία **Provalis Research©**[103] και είναι ένα πακέτο λογισμικού για την ανάλυση δεδομένων και την κωδικοποίηση, ανάκτηση, ανάλυση μεγάλων συλλογών από έγγραφα και εικόνες. Σε συνδυασμό με το πακέτο λογισμικού **WordStat©**, που παρέχεται από την ίδια εταιρεία, επιτρέπει μια ποσοτική ανάλυση των περιεχομένων και χαρακτηριστικών από ένα μεγάλο πλήθος κειμένων. Η έκδοση των δύο πακέτων λογισμικού που χρησιμοποιήσαμε για τις ανάγκες της παρούσας διπλωματικής εργασίας είναι τα trial editions των **QDA Miner© v.4.1.4** και **WordStat© v.6.1**.

Με το **QDA Miner©** καταχωρήσαμε τα αρχεία δεδομένων που χρησιμοποιήσαμε, τα οποία περιείχαν σε μορφή .XML, τις συγκεντρωμένες βιβλιογραφίες αποτελούμενες από άρθρα και δημοσιεύσεις σχετικά με διάφορους όρους, μέσω αναζήτησης στην διαδικτυακή βιβλιοθήκη του **PubMed©** (**Εικόνα 41**). Από τα αρχεία αυτά, επιλέξαμε να κρατήσουμε μόνο το κείμενο της περίληψης, καθότι αυτό περιείχε τις απαραίτητες πληροφορίες που θέλαμε να επεξεργαστούμε, όπως περιγράψαμε στην μεθοδολογία που βρίσκεται στην Παράγραφο 4.3.1. Το λογισμικό αφού επεξεργαστεί το αρχείο .XML, επιστρέφει το πλήθος των περιλήψεων από το κάθε άρθρο του **PubMed©**, υπό την μορφή περιπτώσεων (**Case #**), το κείμενο των οποίων έχει οριστεί ως η μεταβλητή **ABSTRACTTE** (**Εικόνα 42**). Η επεξεργασία αυτών των δεδομένων γίνεται μέσω του λογισμικού **WordStat©**, όπως περιγράφεται στην παράγραφο 4.3.2, και υλοποιείται, αφού διαλέξουμε από τη γραμμή εργασιών του **QDA Miner©** την επιλογή Analyze → Context Analysis και τα δεδομένα που θέλουμε να υποστούν επεξεργασία (στη συγκεκριμένη περίπτωση η μεταβλητή ABSTRACTTE) (**Εικόνα 43)**. Από το σημείο αυτό και μετά, δεν κάνουμε περαιτέρω χρήση του **QDA Miner©**, παρόλα αυτά το πρόγραμμα πρέπει να παραμείνει ανοιχτό, ενόσω επεξεργαζόμαστε τα δεδομένα με το **WordStat©**.

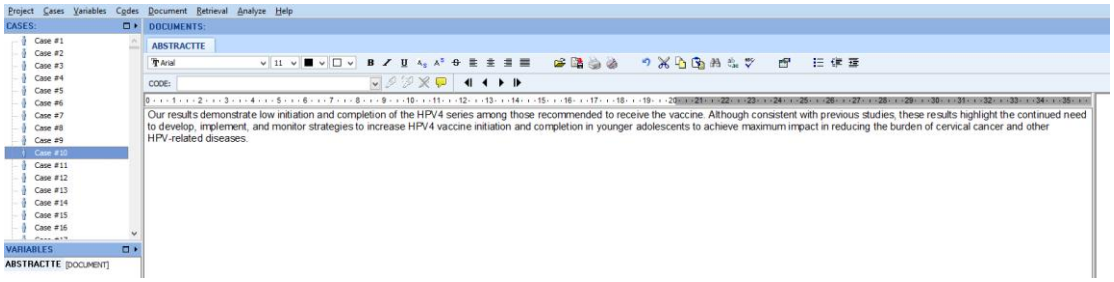
e6e7oncopr10y.xml - Notepad

File Edit Format View Help

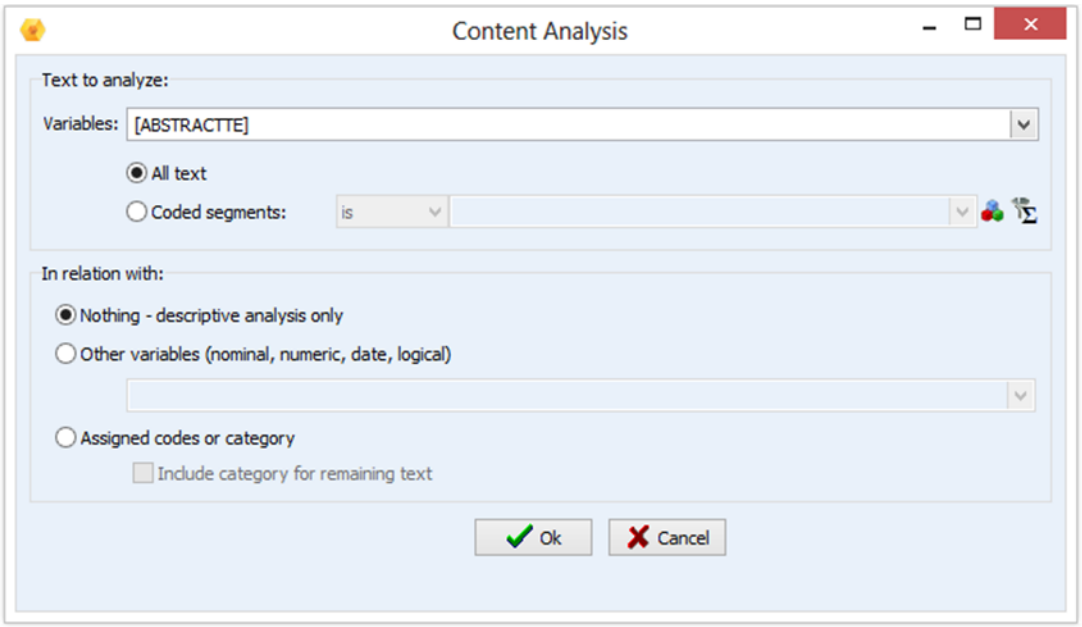
```
<?xml version="1.0"?>
<!DOCTYPE PubmedArticleSet PUBLIC "-//NLM//DTD PubMedArticle, 1st January 2014//EN" "http://www.ncbi.nlm.nih.gov/corehtml/query/DTD/pubmed_140101.dtd">
<PubmedArticleSet>

<PubmedArticle>
    <MedlineCitation Owner="NLM" Status="MEDLINE">
        <PMID Version="1">24199481</PMID>
        <DateCreated>
            <Year>2013</Year>
            <Month>11</Month>
            <Day>08</Day>
        </DateCreated>
        <DateCompleted>
            <Year>2013</Year>
            <Month>12</Month>
            <Day>12</Day>
        </DateCompleted>
        <Article PubModel="Print">
            <Journal>
                <ISSN IssnType="Print">1028-8880</ISSN>
                <JournalIssue CitedMedium="Print">
                    <Volume>16</Volume>
                    <Issue>1</Issue>
                    <PubDate>
                        <Year>2013</Year>
                        <Month>Jan</Month>
                        <Day>1</Day>
                    </PubDate>
                </JournalIssue>
                <Title>Pakistan journal of biological sciences: PJBS</Title>
                <ISOAbbreviation>Pak. J. Biol. Sci.</ISOAbbreviation>
            </Journal>
            <ArticleTitle>Synthesis, characterization and crystal structure of organotin(IV) N-butyl-N-phenyldithiocarbamate compounds and their cytotox
            <Pagination>
                <MedlinePgn>12-21</MedlinePgn>
            </Pagination>
            <Abstract>
                <AbstractText>Organotin complexes are recognized as the biologically active compounds in inducing cancerous cells death at very low dose
the most potent candidates in research related to the new anticancer drugs. In this study, new organotin(IV) N-butyl-N-phenyldithiocarbamate compounds h
```

Εικόνα 41 Αρχείο .xml, πριν την επεξεργασία



Εικόνα 42 Επιφάνεια εργασίας του λογισμικού QDA Miner©, μετά την επιλογή δεδομένων προς επεξεργασία



Εικόνα 43 Επιλογή μεταβλητής ABSTRACTTE, για μετέπειτα επεξεργασία των δεδομένων στο WordStat©

Με την εισαγωγή των δεδομένων στο **WordStat©**, χρειαζόμαστε την πιο βασική μορφή της ανάλυσης περιεχομένου που μπορεί να εκτελέσει το πρόγραμμα και αυτή είναι η ανάλυση συχνότητας εμφάνισης όλων των λέξεων που περιέχονται σε ένα ή περισσότερα πεδία κειμένου από ένα αρχείο δεδομένων. Οι μέθοδοι που ακολουθούνται αφορούν καθαρά την διαδικασία που έχουμε περιγράψει στην ενότητα 1.3.1, σχετικά με την εξόρυξη γνώσης από κείμενα.

Θα εξηγήσουμε τώρα τις δυνατότητες που προσφέρει το συγκεκριμένο πρόγραμμα και θα περιγράψουμε τις διάφορες διαδικασίες που συμμετέχουν σε μια τυπική ανάλυση συχνότητας και πώς αυτές οι διαδικασίες μπορούν να συνδυάζονται για την επίτευξη διαφόρων ειδών ανάλυσης του περιεχομένου ενός αρχείου.

Η κατηγοριοποίηση του **WordStat©** περιλαμβάνει έως πέντε διαδοχικές διαδικασίες:

I. Αποκοπή Καταλήξεων ή Λημματοποίηση

Η διαδικασία της αποκοπής καταλήξεων χρησιμοποιείται για τη μείωση των διαφόρων μορφών μιας λέξης σε ένα περιορισμένο σύνολο λέξεων ή λεκτικών ριζών. Μια τέτοια διαδικασία χρησιμοποιείται συνήθως για τη λημματοποίηση, μια διαδικασία με την οποία όλες οι λέξεις που βρίσκονται σε πληθυντικό αριθμό μετατρέπονται σε ενικό αριθμό και ο αόριστος χρόνος των ρημάτων αντικαθίσταται από τον ενεστώτα. Η διαδικασία αυτή μπορεί επίσης να χρησιμοποιηθεί για αποκοπή καταλήξεων όρων που προέρχονται από ουσιαστικά, ρήματα, επίθετα και επιρρήματα του ίδιου λεκτικού κορμού.

II. Διαδικασία Αποκλεισμού

Η διαδικασία αποκλεισμού μπορεί να εφαρμοστεί για να αφαιρέσει τις λέξεις που δεν θέλουμε να συμπεριληφθούν στην ανάλυση του περιεχομένου των κειμένων. Αυτή η διαδικασία απαιτεί τον προσδιορισμό ενός καταλόγου αποκλεισμού (**Exclusion list**). Μια τέτοια διαδικασία χρησιμοποιείται κυρίως για να καταργηθούν λέξεις με μικρή σημασιολογική αξία όπως αντωνυμίες, άρθρα και σύνδεσμοι, αλλά μπορεί επίσης να χρησιμοποιηθεί για την αφαίρεση κάποιων λέξεων που χρησιμοποιούνται πολύ συχνά ή που έχουν πολύ μικρή διακριτική αξία.

III. Διαδικασία Καταχώρισης και Κατηγοριοποίησης

Η διαδικασία της καταχώρισης και της κατηγοριοποίησης επιτρέπει στον χρήστη να αλλάξει συγκεκριμένες λέξεις, μοτίβα λέξεων ή φράσεις σε άλλες λέξεις, λέξεις-κλειδιά ή κατηγορίες λέξεων και να εξάγει μια λίστα από συγκεκριμένες λέξεις ή όρους. Αυτή η διαδικασία απαιτεί τον προσδιορισμό ενός καταχωρημένου λεξικού, το οποίο μπορεί να χρησιμοποιηθεί για να αφαιρέσει τις διάφορες παραλλαγές μιας λέξης και να τις αντιμετωπίσει ως μία λέξη. Το λεξικό μπορεί επίσης να χρησιμοποιηθεί ως "λεκτικός θησαυρός" (Thesaurus) και να εκτελεί αυτόματη κωδικοποίηση σε κατηγορίες ή σε λεκτικές έννοιες.

IV. Προσθήκη Συχνά Εμφανιζόμενων Λέξεων

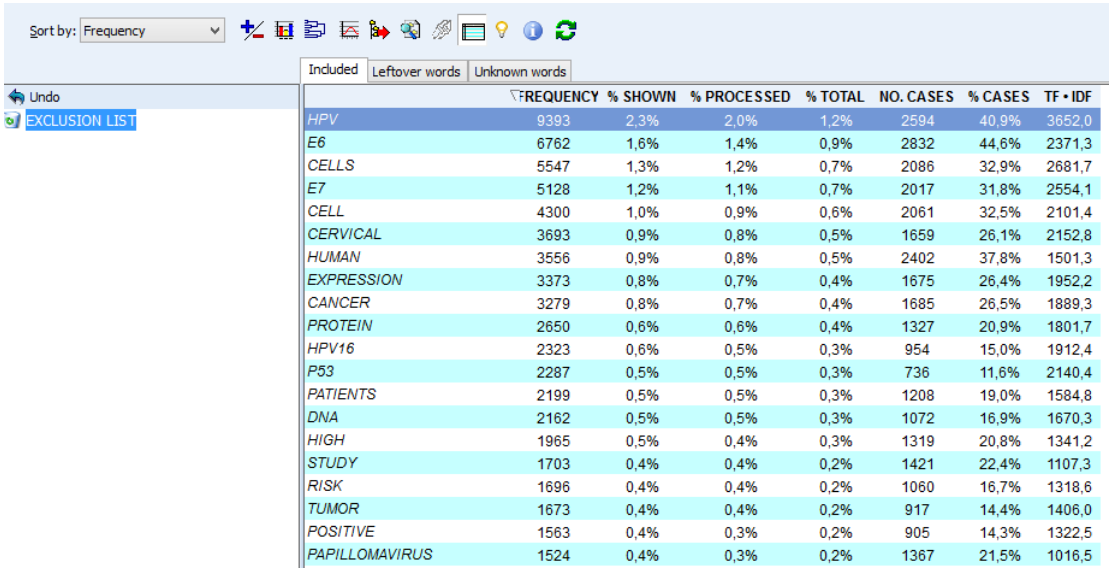
Η επόμενη διαδικασία είναι η εφαρμογή ενός κριτηρίου συχνότητας που χρησιμοποιείται για να προσθέσει στις συμπεριλαμβανόμενες λέξεις ή κατηγορίες λέξεων, τις λέξεις που χρησιμοποιούνται περισσότερο από ένα συγκεκριμένο αριθμό φορών ή που βρίσκονται σε περισσότερες από ένα συγκεκριμένο αριθμό περιπτώσεις. Όταν χρησιμοποιείται ένα λεξικό, αυτή η επιλογή θα προσαρτήσει σε αυτή τη λίστα των περιλαμβανομένων όρων ή κατηγοριών, όλες τις λέξεις που πληρούν το κριτήριο της ελάχιστης συχνότητας. Αν δεν χρησιμοποιείται λεξικό ή κάποια μορφή κατηγοριοποίησης, όλες οι λέξεις που ανταποκρίνονται σε αυτή την ελάχιστη απαίτηση, και που δεν έχουν αποκλειστεί (Διαδικασία ΙΙ), θα προστεθούν στον κατάλογο των τελικών λέξεων/κατηγοριών.

V. <u>Αφαίρεση Λέξεων ή Κατηγοριών</u>

Όταν αυτή η διαδικασία έχει επιλεγεί, όλες οι λέξεις ή κατηγορίες που δεν ικανοποιούν μια ελάχιστη συχνότητα ή κάποιο κριτήριο εμφάνισης περιπτώσεων, θα αφαιρεθούν από τη λίστα με τις τελικές λέξεις/κατηγορίες. Αυτή η διαδικασία μπορεί να συνδυαστεί με τη διαδικασία ένταξης και κατηγοριοποίησης, για να αφαιρέσει τις σπάνια εμφανιζόμενες κατηγορίες. Μπορεί επίσης να χρησιμοποιηθεί σε συνδυασμό με ένα κριτήριο προσθήκης (Διαδικασία IV), ώστε να παρέχει ένα σύνθετο κριτήριο ένταξης, που να περιλαμβάνει τόσο μια ελάχιστη συχνότητα εμφάνισης μιας λέξης-κλειδί, όσο και ενός ελάχιστου αριθμού περιπτώσεων που εμφανίζεται.

Δεδομένου ότι η εφαρμογή της κάθε διαδικασίας είναι προαιρετική, είναι δυνατοί αρκετοί συνδυασμοί και κάθε συνδυασμός επιτρέπει στον ερευνητή να εκτελέσει διαφορετικούς τύπους ανάλυσης περιεχομένων. Στην παρούσα εργασία, οι διαδικασίες που χρησιμοποιήσαμε καθ' όλη την έκταση της εργασίας, ήταν κυρίως οι Διαδικασίες II, IV και V, όπου αυτές κρίθηκαν απαραίτητο να εφαρμοστούν.

Στη συνέχεια περιγράφουμε την καρτέλα εμφάνισης των αποτελεσμάτων, μετά από την διεκπεραίωση των παραπάνω διαδικασιών για την επεξεργασία των δεδομένων στο **WordStat©** (**Εικόνα 44**).



	FREQUENCY	% SHOWN	% PROCESSED	% TOTAL	NO. CASES	% CASES	TF • IDF
HPV	9393	2,3%	2,0%	1,2%	2594	40,9%	3652,0
E6	6762	1,6%	1,4%	0,9%	2832	44,6%	2371,3
CELLS	5547	1,3%	1,2%	0,7%	2086	32,9%	2681,7
E7	5128	1,2%	1,1%	0,7%	2017	31,8%	2554,1
CELL	4300	1,0%	0,9%	0,6%	2061	32,5%	2101,4
CERVICAL	3693	0,9%	0,8%	0,5%	1659	26,1%	2152,8
HUMAN	3556	0,9%	0,8%	0,5%	2402	37,8%	1501,3
EXPRESSION	3373	0,8%	0,7%	0,4%	1675	26,4%	1952,2
CANCER	3279	0,8%	0,7%	0,4%	1685	26,5%	1889,3
PROTEIN	2650	0,6%	0,6%	0,4%	1327	20,9%	1801,7
HPV16	2323	0,6%	0,5%	0,3%	954	15,0%	1912,4
P53	2287	0,5%	0,5%	0,3%	736	11,6%	2140,4
PATIENTS	2199	0,5%	0,5%	0,3%	1208	19,0%	1584,8
DNA	2162	0,5%	0,5%	0,3%	1072	16,9%	1670,3
HIGH	1965	0,5%	0,4%	0,3%	1319	20,8%	1341,2
STUDY	1703	0,4%	0,4%	0,2%	1421	22,4%	1107,3
RISK	1696	0,4%	0,4%	0,2%	1060	16,7%	1318,6
TUMOR	1673	0,4%	0,4%	0,2%	917	14,4%	1406,0
POSITIVE	1563	0,4%	0,3%	0,2%	905	14,3%	1322,5
PAPILLOMAVIRUS	1524	0,4%	0,3%	0,2%	1367	21,5%	1016,5

Εικόνα 44 Καρτέλα Εμφάνισης Αποτελεσμάτων έπειτα από Επεξεργασία Δεδομένων στο WordStat©

Η καρτέλα των αποτελεσμάτων χρησιμοποιείται για την εμφάνιση ενός πίνακα συχνότητας λέξεων ή ονομάτων κατηγοριών. Η καρτέλα αυτή μπορεί να χρησιμοποιηθεί για μια μονοπαραγοντική ανάλυση συχνότητας εμφάνισης λέξεων και επίσης για να τροποποιηθεί κάποιο λεξικό, αν αυτό έχει χρησιμοποιηθεί.

Από προεπιλογή, ο πίνακας δείχνει τις εμφανιζόμενες λέξεις σε φθίνουσα σειρά συχνότητας. Ο πίνακας περιλαμβάνει τα ακόλουθα στατιστικά στοιχεία:

- **Frequency**: αριθμός των εμφανίσεων μιας λέξης ή κατηγορίας.
- **% Shown**: ποσοστό με βάση το συνολικό αριθμό των λέξεων που εμφανίζονται στον πίνακα
- **% Processed** :ποσοστό με βάση το συνολικό αριθμό των λέξεων που επεξεργάσθηκαν κατά τη διάρκεια της ανάλυσης.
- **% Total** : ποσοστό με βάση τον συνολικό αριθμό των λέξεων εκτός εκείνων που εξαιρέθηκαν από τη λίστα.
- **No Cases** : ορισμένες περιπτώσεις όπου εμφανίζεται η συγκεκριμένη λέξη-κλειδί.
- **% Cases** : ποσοστό των περιπτώσεων όπου η συγκεκριμένη λέξη-κλειδί εμφανίζεται.
- **TF*IDF** : η συχνότητα του όρου σταθμισμένη με αντίστροφη συχνότητα του εγγράφου. Η στάθμιση αυτή είναι με βάση την υπόθεση ότι όσο πιο συχνά ένας όρος εμφανίζεται σε ένα έγγραφο, τόσο περισσότερο εκπροσωπεί το περιεχόμενο του εγγράφου ωστόσο όσο περισσότερα τα έγγραφα στα οποία ο όρος εμφανίζεται, τόσο μειώνεται η διακριτική του ικανότητα.

Όπως παρουσιάστηκε και στην ενότητα 1.3.3.1, το σημαντικότερο μέτρο συχνότητας για να αναλύσουμε τη σημαντικότητα και τη διακριτική ικανότητα ενός όρου, είναι ο συντελεστής **TF*IDF**, βάσει του οποίου έγινε και η κατάταξη των όρων, στις περιπτώσεις ιατρικών υποθέσεων που μελετήσαμε.

Βιβλιογραφία

[1] Armstrong EP (April 2010). "Prophylaxis of Cervical Cancer and Related Cervical Disease: A Review of the Cost-Effectiveness of Vaccination against Oncogenic HPV Types". *Journal of Managed Care Pharmacy* **16** (3): 217–30.

[2] "GLOBOCAN 2002 database: summary table by cancer". Archived from the original on 2008-06-16.

[3] http://en.wikipedia.org/wiki/SMART_Information_Retrieval_System

[4] http://wordnet.princeton.edu/

[5] Text Mining for Biology and Biomedicine, Ananiadou and McNaught, 2006

[6] Harnad, Brody, Vallieres, Carr, Hitchcock, Gingras, Oppenheim, Stamerjohanns, and Hilf, 2004

[7] Cohen, and Hunter 2008; Hirschman, Park, Tsujii, and Wong, 2002; (McNaught and Black, 2006) (Jensen, Saric, and Bork, 2006; Hearst, 1999).

[8] http://www.ncbi.nlm.nih.gov/pubmed

[9] Biomedical text mining and its applications in cancer research. Zhu F, Patumcharoenpol P, Zhang C, Yang Y, Chan J, Meechai A, Vongsangnak W, Shen B. J Biomed Inform. 2013 Apr;46(2):200-11. doi: 10.1016/j.jbi.2012.10.007. Epub 2012 Nov 15.

[10] Maier D, Kalus W, Wolff M, Kalko SG, Roca J, Marin de Mas I, et al. Knowledge management for systems biology a general and visually driven framework applied to translational medicine. BMC Syst Biol 2011; 5:38.

[11] Ai J, Smith B, Wong DT. Saliva Ontology: an ontology-based framework for a Salivaomics Knowledge Base. BMC Bioinformatics 2010; 11:302.

[12] Matos S, Arrais JP, Maia-Rodrigues J, Oliveira JL. Concept-based query expansion for retrieving gene related publications from MEDLINE. BMC Bioinformatics 2010; 11:212.

[13] Kuo CJ, Ling MH, Lin KT, Hsu CN. BIOADI: a machine learning approach to identifying abbreviations and definitions in biological literature. BMC Bioinformatics 2009; 10(Suppl. 15):S7.

[14] H. Ao and T. Takagi. Alice: an algorithm to extract abbreviations from medline. J.Am. Med. Inform. Assoc., 12:576–586, 2005.

[15] http://www.ganswers.net/GeneTUKit/demo.jsp

[16] Zhiyong Lu H-YK, Chih-Hsuan Wei, Minlie Huang, Jingchen Liu, Cheng-Ju Kuo, Chun-Nan Hsu, et al. The gene normalization task in BioCreative III. BMC Bioinformatics 2011;12

[17] Ben Abacha A, Zweigenbaum P. Automatic extraction of semantic relations between medical entities: a rule based approach. J Biomed Semantics 2011; 2(Suppl. 5):S4.

[18] http://metamap.nlm.nih.gov/

[19] Eskin E, Agichtein E. Combining text mining and sequence analysis to discover protein functional regions. Pac Symp Biocomput 2004:288–99

[20] Tsai FS. Text mining and visualisation of Protein–Protein Interactions. Int J Comput Biol Drug Des 2011;4:239–44

[21] Agarwal S, Liu F, Yu H. Simple and efficient machine learning frameworks for identifying protein–protein interaction relevant articles and experimental methods used to study the interactions. BMC Bioinformatics 2011; 12(Suppl. 8):S10.

[22] http://zope.bioinfo.cnio.es/plan2l/plan2l.html

[23] Korhonen A, Seaghdha DO, Silins I, Sun L, Hogberg J, Stenius U. Text mining for literature review and knowledge discovery in cancer risk assessment and research. PLoS One 2012; 7:e33427.

[24] Nam S, Park T. Pathway-based evaluation in early onset colorectal cancer suggests focal adhesion and immunosuppression along with epithelial–mesenchymal transition. PLoS One 2012; 7:e31685.

[25] Urzua U, Owens GA, Zhang GM, Cherry JM, Sharp JJ, Munroe DJ. Tumor and reproductive traits are linked by RNA metabolism genes in the mouse ovary: a transcriptome-phenotype association analysis. BMC Genomics 2010;11 (Suppl. 5):S1.

[26] Li J, Zhu X, Chen JY. Building disease-specific drug–protein connectivity maps from molecular interaction networks and PubMed abstracts. PLoS Comput Biol 2009;5:e1000450

[27] Topinka CM, Shyu CR. Predicting cancer interaction networks using text mining and structure understanding. In: AMIA annu symp proc; 2006. p.1123.

[28] http://www.nlm.nih.gov/pubs/factsheets/medline.html

[29] http://www.textpresso.org/

[30] http://www.gopubmed.org/web/gopubmed/

[31] http://www.ihop-net.org/UniPub/iHOP/

[32] http://www.ensembl.org/index.html
[33] http://biotextquest.ucy.ac.cy/cgi-bin/textQuest/textQuest.cgi
[34] http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html
[35] http://ibmi3.mf.uni-lj.si/bitola/
[36] http://pages.cs.wisc.edu/~bsettles/abner/
[37] http://www.nactem.ac.uk/GENIA/tagger/
[38] http://alias-i.com/lingpipe/
[39] https://www.sics.se/search/content/humle%20projects%20prothalt
[40] http://www.nactem.ac.uk/GENIA/tagger/
[41] http://www.ebi.ac.uk/Rebholz-srv/BioLexicon/biolexicon.html
[42] ftp://ftp.ncbi.nlm.nih.gov/pub/tanabe
[43] http://www.geneontology.org/
[44] http://bcms.bioinfo.cnio.es/
[45] http://www.chilibot.net/
[46] http://wilab.inha.ac.kr/HPID/
[47] http://www.hprd.org/
[48] http://www.ihop-net.org/UniPub/iHOP/
[49] http://www.ebi.ac.uk/intact/
[50]http://www.elsevier.com/about/mission/innovative-tools/elsevier-acquires-ariadne-genomics, provider-of-pathway-analysis-tools-and-semantic-technologies-for-life-science-researchers
[51] http://www.pubgene.org/
[52] http://www.reactome.org/ReactomeGWT/entrypoint.html
[53] http://mars.cs.utu.fi/BioInfer/
[54] http://www.ncbi.nlm.nih.gov/projects/RefSeq/HIVInteractions/
[55] http://genome.jouy.inra.fr/~cnedelle/Docs/LLL-challenge-05.pdf
[56] http://bionlp-corpora.sourceforge.net/picorpus/
[57] http://icb.med.cornell.edu/services/pdz/start
[58] http://string-db.org/
[59] http://www.biocreative.org/
[60] http://www.biotext.com.au/
[61] http://www.nactem.ac.uk/genia/
[62] Blagosklonny MV, Pardee AB: Unearthing the gems. Nature 2002, 416:373.
[63] http://www.fao.org/docrep/w3241e/w3241e02.htm#step 2: hypothesis generation
[64] Swanson DR: Raynaud's syndrome and undiscovered public knowledge. In Perspectives in Biology and Medicine. Volume 30. John Hopkins University Press; 1984:(1):7-18.
[65] Bekhuis T: Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. Biomed Digit Libr 2006, 3:2
[66] http://arrowsmith.psych.uic.edu/arrowsmith_uic/tools.html
[67] http://gateway.ovid.com/
[68] Gordon MD, Lindsay RK: Toward discovery support systems: a replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. Journal of the American Society for Information Science 1996, 47(2):116-128
[69] Weeber M, Klein H, Aronson AR, Mork JG, de Jong-van den Berg LTW, Vos R: Text-based discovery in biomedicine: the architecture of the DAD-system. In *Proceedings of the AMIA Annual Fall Symposium* Edited by: Overhage JM. Philadelphia, Hanley & Belfus; 2000:903-907.
[70] http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html
[71] http://www.nlm.nih.gov/
[72] http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html
[73] Weeber M, Vos R, Klein H, de Jong-van den Berg LTW, Aronson AR, Molema G: Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. Journal of the American Medical Informatics Association 2003, 10:252-259.
[74] Stegmann J, Grohmann G: Hypothesis generation guided by coword clustering. Scientometrics 2003, 56(1):111-135.
[75] Chen C: Mapping Scientific Frontiers: The Quest for Knowledge Visualization. London, Springer-Verlag; 2003.

[76] Srinivasan P: Text mining: generating hypotheses from Medline. Journal of American Society for Information Science and Technology 2004, 55(5):396-413.
[77] Srinivasan P, Libbus B: Mining MEDLINE for implicit links between dietary substances and diseases. Bioinformatics 2004, 20(Suppl. 1):i290-i296.
[78] Pratt W, Yildiz MY: Capturing connections across the biomedical literature. Proceedings of the 2nd International Conference on Knowledge Capture; Sanibel Island, FL, USA ACM; 2003, 105-112
[79] Katukuri et al: Hypotheses generation as supervised link discovery with automated class labeling on largescale biomedical concept networks. BMC Genomics 2012, 13(Suppl 3):S5
[80] http://www.informatik.uni-trier.de/~ley/db/
[81] Őzgür A, Vu T, Ergan G, Radev DR: Identifying gene-disease associations using centrality on a literature mined gene-interaction network. Bioinformatics 2008, 24(13):i277-i285.
[82] Ceci, F., Pietrobon, R. & Goncalves, A. L. Turning text into research networks: information retrieval and computational ontologies in the creation of scientific databases. PLoS ONE 7, e27499 (2012).
[83] Pesquita, C. et al. Metrics for GO based protein semantic similarity: a systematic evaluation. BMC Bioinformatics 9 (Suppl. 5), S4 (2008).
[84] Coulet, A., Shah, N. H., Garten, Y., Musen, M. & Altman, R. B. Using text to build semantic networks for pharmacogenomics. J. Biomed. Informat. 43, 1009–1019 (2010).
[85] Percha, B., Garten, Y. & Altman, R. B. Discovery and explanation of drug-drug interactions via text mining. Pacific Symp. Biocomput. 2012, 410–421 (2012).
[86] Campillos, M., Kuhn, M., Gavin, A. -C., Jensen, L. J. & Bork, P. Drug target identification using side-effect similarity. Science 321, 263–266 (2008).
[87] http://swan.mindinformatics.org/
[88] Gao Y, Kinoshita J, Wu E, et al: SWAN: A Distributed Knowledge Infrastructure for Alzheimer Disease Research. J. Of Web Semantics 2006, 4(3):222-228.
[89] The OBI Consortium:[http://obi-ontology.org]
[90] The ART project:[http://www.jisc.ac.uk/whatwedo/programmes/reppres/tools/art.aspx]
[91] Soldatova LN, King RD: An Ontology of Scientific Experiments. J R Soc Interface 2006, 3(11):795-803.
[92] The HyBrow project:[http://www.hybrow.org].
[93] Racunas SA, Shah NH, Albert I, Fedoroff NV: Hybrow: a prototype system for computer-aided hypothesis evaluation. Bioinformatics 2004, 20(Suppl. 1).
[94] The Yeast Proteome:[http://www.proteome.com].
[95]Large-Scale Discovery of Scientific Hypotheses project:
[http://arrafunding.uchicago.edu/investigators/rzhetsky_a.shtml].
[96] http://provalisresearch.com/products/qualitative-data-analysis-software/
[97] http://provalisresearch.com/products/content-analysis-software/
[98] Γυναικολογία και μαιευτική της νεαρής ηλικίας, ISBN: 9789608122819, Συγγραφέας: Γεώργιος Κ. Κρεατσάς, Εκδότης: Ιατρικές Εκδόσεις Π. Χ. Πασχαλίδης
[99] The natural history of human papillomavirus infection as measured by repeated DNA testing in adolescent and young women. Moscicki AB, Shiboski S, Broering J, Powell K, Clayton L, Jay N, Darragh TM, Brescia R, Kanowitz S, Miller SB, Stone J, Hanson E, Palefsky J., J Pediatr. 1998 Feb;132(2):277-84.
[100] Γυναικολογία ΙΙΙ, ISBN: 9607398653, Συγγραφέας: Emil Novak, Εκδότης: Ιατρικές Εκδόσεις Π. Χ. Πασχαλίδης
[101] http://www.webmd.com/lung-cancer/news/20130410/study-hints-of-links-between-hpv-and-lung-cancer
[102] http://link.springer.com/chapter/10.1007%2F978-1-4419-1472-9
[103] http://provalisresearch.com/