ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Υλοποίηση μηχανισμού κατάταξης δημοσιεύσεων σχετικών με βιομόρια micro-RNA

Διπλωματικη Εργασια

της

ΒΑΣΙΛΙΚΗΣ Α. ΒΛΑΧΟΚΥΡΙΑΚΟΥ

Επιβλέπων: Ιωάννης Βασιλείου
Καθηγητής Ε.Μ.Π.

Αθήνα, Δεκέμβριος 2015

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Υλοποίηση μηχανισμού κατάταξης δημοσιεύσεων σχετικών με βιομόρια micro-RNA

Διπλωματικη Εργασια

της

ΒΑΣΙΛΙΚΗΣ Α. ΒΛΑΧΟΚΥΡΙΑΚΟΥ

Επιβλέπων: Ιωάννης Βασιλείου
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή τη 18η Δεμεκεμβρίου 2015.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Ιωάννης Βασιλείου	Νεκτάριος Κοζύρης	Θεόδωρος Δαλαμάνγκας
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Επιστ. Συνεργάτης

Αθήνα, Δεκέμβριος 2015

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

(Υπογραφή)

.....

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

Περίληψη

Η μελέτη των βιομορίων που συμμετέχουν στους μηχανισμούς της ζωής (πχ DNA, πρωτεΐνες, μόρια microRNA κτλ), είναι απαραίτητη στην έρευνα για τη θεραπεία γενετικών ασθενειών. Οι σχετικές με τα βιομόρια εργασίες, δημοσιεύονται σε επιστημονικά περιοδικά και συνέδρια, και μαζί με τα μεταδεδομένα που αφορούν στο περιεχόμενό τους, καταγράφονται σε βάσεις δεδομένων που προσφέρουν δυνατότητες αναζήτησης μέσω Διαδικτύου.

Σε προηγούμενες εργασίες, έχει κατασκευαστεί η εφαρμογή Diana Mirpub, μια βάση δεδομένων που καταγράφει τις επιστημονικές δημοσιεύσεις σχετικές με τα βιομόρια microRNA και μια διαδικτυακή εφαρμογή που παρέχει προηγμένες υπηρεσίες αναζήτησης πάνω στα δεδομένα που αποθηκεύονται στη συγκεκριμένη βάση. Οι εργασίες που επιστρέφονται ως αποτέλεσμα μιας αναζήτησης, παρουσιάζονται με βάση τη χρονολογία δημοσίευσης. Στόχος της παρούσας εργασίας είναι η υλοποίηση ενός μηχανισμού κατάταξης για την παρουσίαση των αποτελεσμάτων της εφαρμογής με βάση τη σημαντικότητα των εργασιών.

Αντίστοιχοι μηχανισμοί, που χρησιμοποιούνται από τις μηχανές αναζήτησης για την κατάταξη των ιστοσελίδων στο Διαδίκτυο, αδυνατούν να κατατάξουν σωστά τις νέες δημοσιεύσεις, εξαιτίας των διαφορών που παρουσιάζουν οι γράφοι αναφορών επιστημονικών δημοσιεύσεων σε σχέση με τους γράφους υπερσυνδέσεων μεταξύ ιστοσελίδων. Μελετήθηκαν, λοιπόν, αλγόριθμοι κατάταξης δημοσιεύσεων που προτείνονται από τη βιβλιογραφία και αναπτύχθηκαν προγράμματα που υλοποιούν τους αλγόριθμους αυτούς. Τέλος, υλοποιήθηκε στο Diana MirPub η λειτουργικότητα που παρέχει στον χρήστη τη δυνατότητα επιλογής μηχανισμού κατάταξης όπου ενσωματώθηκαν οι αλγόριθμοι που αξιολογήθηκαν ως βέλτιστοι.

Λέξεις Κλειδιά

δίκτυο δημοσιεύσεων, παραπομπές, περιοδικά, μηχανισμός κατάταξης

Abstract

The study of biomolecules involved in different mechanisms of life (DNA, proteins, molecules microRNA etc.), is necessary for researchers to understand and cure genetic diseases. Papers related to these biomolecules, are published in scientific journals and conferences, and, along with the metadata associated with their content, are recorded in databases that offer searching abilities online.

In previous work, the application Diana MirPub has been developed which is a database that contains scientific papers relevant to microRNA biomolecules, combined with a web application that provides advanced search services on stored data. The scientific papers that form the result set of a search query, are ranked according to their date of publication. The objective of this thesis is to implement a ranking mechanism for presenting the results of the application based on the importance of each work.

Similar mechanisms, used by search engines to rank websites, are unable to evaluate correctly the importance of new publications, due to the differences found in graphs of citations between publications, to those of hyperlinks between websites. Therefore, publication ranking algorithms proposed in relevant bibliography have been studied and programs that implement them have been developed. Finally the implemented algorithms evaluated as optimal were incorporated on Diana MirPub. A new functionality was added to the application enabling a user to display search results, based on one of these algorithms.

Keywords

publications graph, citations, journals, ranking mechanism

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

Στο κεφάλαιο αυτό, παρουσιάζεται το αντικείμενο συγκεκριμένης διπλωματικής: η μελέτη των αλγορίθμων και η υλοποίηση ενός μηχανισμού κατάταξης των επιστημονικών εργασιών που περιέχονται στη βάση δεδομένων της εφαρμογής Diana MirPub.

1.1 Αντικείμενο διπλωματικής

Το σύνολο της γενετικής πληροφορίας ενός οργανισμού κωδικοποιείται σε ακολουθίες DNA, που ονομάζονται γονίδια. Το κύτταρο «διαβάζει» τη γενετική πληροφορία που κωδικοποιούν τα γονίδια και, με βάση αυτή, παράγει πρωτεΐνες, θέτοντας έτσι σε εφαρμογή τους μηχανισμούς της ζωής. Δυσλειτουργίες κατά την παραγωγή πρωτεϊνών μπορούν να δημιουργήσουν προβλήματα στους μηχανισμούς αυτούς. Τέτοιες δυσλειτουργίες αποτελούν την αιτία πολλών γενετικών ασθενειών. Τα μόρια microRNA, τα οποία λειτουργούν ως ρυθμιστές της παραγωγής πρωτεϊνών, μπορούν να βοηθήσουν στην κατανόηση και πιθανώς την αντιμετώπιση τέτοιων ασθενειών.

Στο Ινστιτούτο Πληροφοριακών Συστημάτων του ΕΚ «Αθηνά» και σε συνεργασία με το ΕΚ Βιοϊατρικής «Αλέξανδρος Φλέμινγκ» έχει αναπτυχθεί η εφαρμογή Diana Mirpub, μια βάση δεδομένων για την καταγραφή επιστημονικών δημοσιεύσεων σχετικών με τα βιομόρια microRNA και μια διαδικτυακή εφαρμογή που προσφέρει προηγμένες υπηρεσίες αναζήτησης πάνω στα δεδομένα που αποθηκεύονται στη συγκεκριμένη βάση. Η εφαρμογή λαμβάνει υπόψη τόσο τα διαθέσιμα μέρη των κειμένων των εργασιών όσο και μεταδεδομένα που τα συνοδεύουν (π.χ. πληροφορίες για το περιεχόμενό τους που έχουν καταγραφεί είτε από τους συγγραφείς των εργασιών είτε από επιμελητές που τις έχουν μελετήσει) για να προσφέρει στους χρήστες της τον πληρέστερο κατάλογο δημοσιεύσεων που είναι σχετικές με τις αναζητήσεις τους.

Στην παρούσα φάση, τα αποτελέσματα των αναζητήσεων επιστρέφονται καταταγμένα με βάση την ημερομηνία δημοσίευσης ξεκινώντας από τις πιο πρόσφατες εργασίες και καταλήγοντας στις πιο παλιές. Όμως, ο συγκεκριμένος τρόπος παρουσίασης δεν είναι ιδανικός, καθώς μπορεί σημαντικές δημοσιεύσεις να κρύβονται αρκετά χαμηλά στη λίστα των αποτελεσμάτων ενώ εργασίες ήσσονος σημασίας να εμφανίζονται ως πρώτα αποτελέσματα. Είναι απαραίτητο λοιπόν να χρησιμοποιηθούν κάποια ποιοτικά κριτήρια που θα δίνουν μεγαλύτερη προτεραιότητα

εμφάνισης στις πιο σημαντικές εργασίες. Επομένως, αντικείμενο της διπλωματικής είναι η υλοποίηση ενός μηχανισμού κατάταξης (ranking) για την παρουσίαση των αποτελεσμάτων της εφαρμογής με βάση διάφορα κριτήρια (π.χ. σημαντικότητα περιοδικού, σημαντικότητα εργασίας), έτσι ώστε οι πιο σημαντικές εργασίες να επιστρέφονται στις πρώτες θέσεις της λίστας αποτελεσμάτων.

1.2 Οργάνωση του τόμου

Η οργάνωση του κειμένου της διπλωματικής εργασίας έχει πραγματοποιηθεί ως εξής:

Στο Κεφάλαιο 2 παρουσιάζεται η δομή του διαδικτυακού γράφου και αναλύεται η λειτουργία του πιο διαδεδομένου αλγόριθμου κατάταξης των ιστοσελίδων, του PageRank, καθώς και μίας παραλλαγής του, που λαμβάνει υπόψη την αδυναμία του PageRank να αξιολογήσει σωστά τις πρόσφατες ιστοσελίδες. Στη συνέχεια, περιγράφεται η δομή του γράφου των αναφορών μεταξύ των δημοσιεύσεων και ο τρόπος με τον οποίο πραγματοποιείται η κατάταξη των δημοσιεύσεων στις αντίστοιχες μηχανές αναζήτησης. Γι' αυτό το λόγο, παρουσιάζονται δύο μηχανισμοί αξιολόγησης περιοδικών που χρησιμοποιούνται ευρέως, ο Impact Factor και ο EigenFactor, και μελετώνται σχετικές εργασίες στις οποίες έχουν προταθεί μηχανισμοί κατάταξης δημοσιεύσεων. Τέλος, περιγράφεται η εφαρμογή Diana MirPub, που καλούμαστε να επεκτείνουμε με τα αποτελέσματα της παρούσας διπλωματικής.

Στο Κεφάλαιο 3 παρουσιάζονται όλοι οι αλγόριθμοι που υλοποιήθηκαν με βάση τις παραπάνω εργασίες, καθώς και τα εργαλεία που χρησιμοποιήθηκαν για την ανάπτυξη των αντίστοιχων προγραμμάτων.

Στο Κεφάλαιο 4 αξιολογούνται οι υλοποιημένοι μηχανισμοί κατάταξης. Αρχικά, αναφέρονται οι μετρικές που χρησιμοποιήθηκαν και περιγράφεται η μέθοδος αξιολόγησης, ενώ ακολουθούν τα αποτελέσματα και τα συμπεράσματα της εφαρμογής της στις εξόδους των αλγόριθμων. Με βάση την αξιολόγηση, επιλέγονται οι καλύτεροι μηχανισμοί κατάταξης ώστε να ενσωματωθούν στο σύστημα.

Στο Κεφάλαιο 5 περιγράφεται η υλοποίηση της λειτουργικότητας ώστε οι δημοσιεύσεις της εφαρμογής Diana Mirpub να κατατάσσονται με βάση τους επιλεγμένους αλγόριθμους. Αρχικά δίνονται κάποιες λεπτομέρειες για τη δομή της εφαρμογής και τα σημεία που χρειάστηκε να επέμβουμε για την ανάπτυξη της ζητούμενης λειτουργικότητας, καθώς και τα προγραμματιστικά εργαλεία που αυτή χρησιμοποιεί. Τέλος, περιγράφεται η υλοποιημένη λειτουργικότητα και πραγματοποιείται επίδειξή της.

Στο Κεφάλαιο 6 προτείνονται μελλοντικές επεκτάσεις όσων αναπτύχθηκαν.

Κεφάλαιο 2

Θεωρητικό υπόβαθρο και σχετικές εργασίες

Στο κεφάλαιο αυτό διατυπώνεται το θεωρητικό υπόβαθρο σχετικά με τους αλγόριθμους κατάταξης επιστημονικών εργασιών που εξετάστηκαν στα πλαίσια της διπλωματικής. Αρχικά περιγράφονται τα χαρακτηριστικά των γράφων που παράγονται από τα δίκτυα αναφορών μεταξύ ιστοσελίδων, ο αλγόριθμος Pagerank και μία παραλλαγή του. Στη συνέχεια παρουσιάζονται τα χαρακτηριστικά του γράφου των επιστημονικών εργασιών συγκριτικά με αυτά του γράφου των ιστοσελίδων και ο τρόπος κατάταξης των επιστημονικών εργασίων στις αντίστοιχες μηχανές αναζήτησης. Γι' αυτό το λόγο, παρατίθενται οι μετρικές κατάταξης περιοδικών που χρησιμοποιούνται ευρέως και μελετώνται παραλλαγές του αλγόριθμου Pagerank που έχουν προταθεί στη βιβλιογραφία και εφαρμόζονται για την κατάταξη επιστημονικών εργασιών. Τέλος, περιγράφονται τα χαρακτηριστικά της εφαρμογής Diana Mirpub, η οποία έχει κατασκευαστεί σε προηγούμενες εργασίες.

2.1 Κατάταξη ιστοσελίδων στο Διαδίκτυο

Όπως αναφέρθηκε, στόχος της διπλωματικής είναι η υλοποίηση ενός μηχανισμού κατάταξης των επιστημονικών εργασιών που είναι καταχωρημένες στην εφαρμογή Diana Mirpub. Αντίστοιχοι μηχανισμοί χρησιμοποιούνται από τις μηχανές αναζήτησης γενικού σκοπού για την κατάταξη των ιστοσελίδων στο Διαδίκτυο.

Τα αποτελέσματα που επιστρέφονται από μία μηχανή αναζήτησης με βάση ένα ερώτημα κατατάσσονται σύμφωνα με:

- Τη σχετικότητα της κάθε ιστοσελίδας με το ερώτημα. Οι μηχανές αναζήτησης δημιουργούν ένα «ανεστραμμένο ευρετήριο», το οποίο τυπικά αποθηκεύει για κάθε όρο της συλλογής, το σύνολο των κειμένων (ιστοσελίδων) στα οποία τον συναντάμε καθώς επίσης και κάποιο βάρος, όπως π.χ. το πλήθος των εμφανίσεων του όρου στο συγκεκριμένο κείμενο. Η διαδικασία αυτή ονομάζεται ευρετηρίαση (indexing) και με βάση την πληροφορία που παράγει, αξιολογείται η σχετικότητα των ιστοσελίδων με τον εκάστοτε όρο που θέτει ο χρήστης σαν ερώτημα.

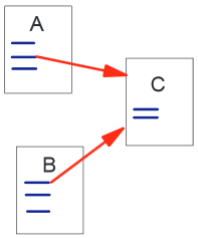
- Το κύρος της ιστοσελίδας. Πέρα από την αξιολόγηση της πληροφορίας που βρίσκεται μέσα στο εκάστοτε κείμενο μίας ιστοσελίδας, μπορούμε να εκμεταλλευτούμε τη δομή του γράφου που απεικονίζει τον τρόπο που συνδέονται οι σελίδες μεταξύ τους στον παγκόσμιο ιστό. Ο αλγόριθμος PageRank χρησιμοποιείται σήμερα από την Google γι' αυτό το λόγο. Η βασική ιδέα είναι ότι κάθε σελίδα στο Διαδίκτυο κατέχει ένα γενικό μέτρο «κύρους» ή «σημαντικότητας», το οποίο είναι ανεξάρτητο από τα επερωτήματα των χρηστών.

Η δομή του διαδικτυακού γράφου

Ο γράφος του Διαδικτύου αποτελείται από ένα μεγάλο πλήθος ιστοσελίδων (κόμβοι) και ένα ακόμα μεγαλύτερο πλήθος συνδέσμων (ακμές) που συνδέουν τις ιστοσελίδες μεταξύ τους. Κάθε ιστοσελίδα περιέχει εξερχόμενους και εισερχόμενους συνδέσμους (οι σύνδεσμοι που αναφέρει και οι σύνδεσμοι που αναφέρονται σε αυτήν αντίστοιχα).

Σε μια δεδομένη χρονική στιγμή δεν μπορούμε να γνωρίζουμε τους εισερχόμενους συνδέσμους μιας ιστοσελίδας, αλλά από τη στιγμή που την κατεβάσουμε τοπικά μπορούμε να γνωρίζουμε τους εξερχόμενους συνδέσμους της. Αυτό συμβαίνει διότι ένας σύνδεσμος εντοπίζεται όταν εξέρχεται από μία σελίδα, εφόσον φορτωθεί και αναγνωσθεί η σελίδα αυτή, μέσω των anchor (<a>) ετικετών.

Η ύπαρξη εξερχόμενων συνδέσμων σε μία σελίδα δεν υποδεικνύει υψηλή ποιότητα για την ίδια τη σελίδα κι ελέγχεται αποκλειστικά από τους διαχειριστές της. Αντίθετα, η ύπαρξη εισερχόμενων συνδέσμων πραγματοποιείται φυσικά, με το αντικειμενικό κριτήριο του ενδιαφέροντος των εξωτερικών ως προς τη σελίδα παραγόντων κι εκφράζει τη δημοτικότητά της. Γενικά όσες ιστοσελίδες διαθέτουν πολλούς εισερχόμενους συνδέσμους μπορούν να θεωρηθούν σημαντικές.



Σχήμα 2.1: Οι σύνδεσμοι των ιστοσελίδων A και B προς τη σελίδα C αποτελούν εισερχόμενες ακμές της τελευταίας.

2.1.1 Μηχανισμοί κατάταξης ιστοσελίδων στο διαδίκτυο

Ο αλγόριθμος PageRank αποτελεί την καρδιά της μηχανής αναζήτησης Google και χρησιμοποιείται για την κατάταξη ιστοσελίδων στο Διαδίκτυο. Ο αλγόριθμος αυτός βασίζεται στην ιδέα ότι όταν μια ιστοσελίδα είναι σημαντική είναι και δημοφιλής, ιδίως σε άλλες σημαντικές σελίδες. Όταν δηλαδή μια σελίδα ενδιαφέρει έναν μεγάλο αριθμό χρηστών, υπάρχουν υπερσύνδεσμοι (εισερχόμενες συνδέσεις) που θα δείχνουν σε αυτήν, κυρίως από σελίδες που και

οι ίδιες έχουν μεγάλο πλήθος εισερχόμενων συνδέσμων. Η ιδέα αυτή υστερεί όταν αναφερόμαστε σε πρόσφατα δημιουργημένες ιστοσελίδες, οι οποίες ακόμα δεν είναι γνωστές. Έτσι, πέρα από τον αλγόριθμο PageRank, στη συγκεκριμένη ενότητα θα παρουσιάσουμε και ένα μοντέλο κατάταξης που αντιμετωπίζει αυτό το πρόβλημα.

2.1.1.1 Ο αλγόριθμος PageRank

Ο αλγόριθμος αυτός χρησιμοποιείται για να μετρηθεί η σχετική σημασία των ιστοσελίδων και να πραγματοποιηθεί μία κατάταξή τους βασισμένη στον διαδικτυακό γράφο. Όπως αναφέραμε παραπάνω, ιστοσελίδες με μεγάλο αριθμό εισερχόμενων συνδέσμων θεωρούνται σημαντικότερες από άλλες που έχουν λιγότερους, επομένως είναι πιο σημαντικές και για τις μηχανές αναζήτησης. Χρησιμοποιώντας όμως μόνο το πλήθος των εισερχόμενων συνδέσμων για να αξιολογήσουμε μια ιστοσελίδα δεν λαμβάνουμε υπόψη την ποιότητα των ιστοσελίδων που φιλοξενούν τους συνδέσμους αυτούς.

Αποδίδοντας λοιπόν μία πιο σφαιρική περιγραφή στον όρο PageRank και τη λειτουργία του, σημειώνουμε ότι: *μία σελίδα κατέχει υψηλό δείκτη θέσης στην ιεραρχία του διαδικτυακού γράφου εάν το άθροισμα των σκορ των εισερχόμενων συνδέσμων προς αυτήν είναι υψηλό, με άλλα λόγια, μία σελίδα κατέχει υψηλό σκορ PageRank εάν το άθροισμα των σκορ PageRank των σελίδων που συνδέουν προς αυτήν είναι υψηλό.*

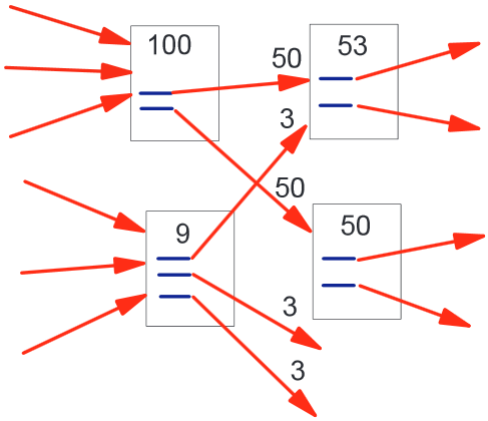
Ορισμός του PageRank

Μια απλουστευμένη μορφή του τύπου που χρησιμοποιείται για τον υπολογισμό του Pagerank είναι η εξής:

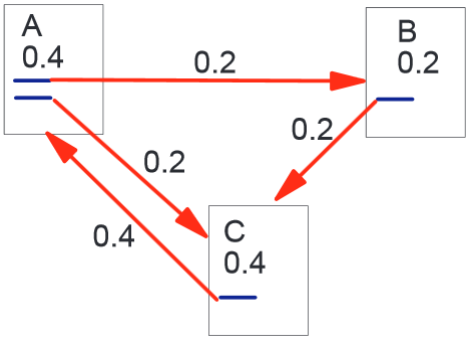
$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v} \tag{2.1}$$

όπου R ο βαθμός Pagerank, B_u το σύνολο των ιστοσελίδων που παραπέμπουν στην ιστοσελίδα u και $N_v = |F_v|$ με F_v το σύνολο των ιστοσελίδων στις οποίες παραπέμπει η ιστοσελίδα v.

Από τον τύπο προκύπτει ότι το σκορ της κατάταξης κάθε ιστοσελίδας διαιρείται και μεταφέρεται στις εξερχόμενες ακμές της όπως παρουσιάζεται στα Σχήματα 2.2 και 2.3. Ο παράγοντας κανονικοποίησης c διορθώνει το πρόβλημα που προκύπτει όταν μία σελίδα λαμβάνει συνδέσμους, αλλά δεν υπάρχουν σελίδες στις οποίες να δείχνει η ίδια, δεν έχει δηλαδή εξερχόμενους συνδέσμους και το βάρος των συνδέσμων που συνεισφέρουν στο σκορ της σελίδας αυτής χάνεται. Έτσι, έχουμε: $c < 1$.



Σχήμα 2.2: Η μετάδοση του σκορ κατάταξης κάθε ιστοσελίδας στις γειτονικές της.



Σχήμα 2.3: Τελικές τιμές του Pagerank.

Ο τύπος είναι αναδρομικός αλλά μπορεί να υπολογιστεί ξεκινώντας από μία δεδομένη τιμή των σκορ κατάταξης και επαναλαμβάνοντας τη διαδικασία μέχρι να οδηγηθούμε σε σύγκλιση.

Εναλλακτικά η παραπάνω απλουστευμένη εκδοχή του Pagerank ορίζεται ως εξής:

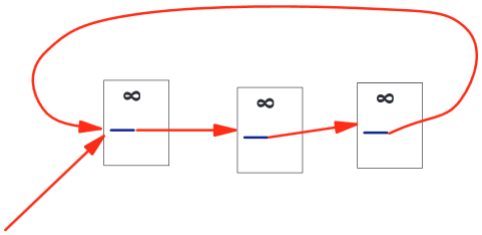
$$\mathbf{R} = c\mathbf{G} \cdot \mathbf{R} \tag{2.2}$$

όπου $\mathbf{G}$ τετραγωνική μήτρα με τις γραμμές u και τις στήλες v να αντιστοιχούν στις ιστοσελίδες του Διαδικτύου και $\mathbf{G}_{u,v} = 1/N_v$, με N_v το πλήθος των εξερχόμενων ακμών του v, όταν υπάρχει ακμή από το v στο u και $\mathbf{G}_{u,v} = 0$ όταν δεν υπάρχει.

Επομένως το $\mathbf{R}$ αποτελεί ιδιοδιάνυσμα της μήτρας $\mathbf{G}$ με ιδιοτιμή τον παράγοντα κανονικοποίησης c. Το $\mathbf{R}$ μπορεί να υπολογιστεί επαναληπτικά ξεκινώντας από ένα μη μηδενικό διάνυσμα.

Ένα πρόβλημα που προκύπτει από τον παραπάνω απλοποιημένο τύπο υπολογισμού του Pagerank είναι το εξής: Έστω δύο ιστοσελίδες που η μία δείχνει στην άλλη χωρίς να έχουν άλλες εξερχόμενες ακμές και μία τρίτη ιστοσελίδα που δείχνει σε μία από αυτές. Κατά τη διάρκεια της επαναληπτικής διαδικασίας, οι ιστοσελίδες αυτές θα λαμβάνουν όλο και μεγαλύτερο σκορ χωρίς όμως να μπορούν να το μεταφέρουν, μιας και δεν έχουν

εξερχόμενες ακμές εκτός του συγκεκριμένου συνόλου. Παράδειγμα τέτοιων ιστοσελίδων παρουσιάζεται στο Σχήμα 2.4.



Σχήμα 2.4: Σύνολο ιστοσελίδων που το σκορ τους αυξάνεται χωρίς να μεταφέρεται σε άλλες ιστοσελίδες.

Για την αντιμετώπιση αυτού του φαινομένου, επινοήθηκε ένα αρχικό διάνυσμα κατάταξης $E(u)$, με τη χρήση του οποίου ορίστηκε ο τύπος υπολογισμού του αλγόριθμου:

Έστω $E(u)$ διάνυσμα που απεικονίζει μια αρχική κατάταξη των ιστοσελίδων. Το σκορ PageRank των ιστοσελίδων, R', ικανοποιεί τον παρακάτω τύπο:

$$R'(u) = c \sum_{v \in B_u} \frac{R'(v)}{N_v} + cE(u) \tag{2.3}$$

έτσι ώστε ο παράγοντας κανονικοποίησης c να μεγιστοποιείται και $\|R'\|_1 = 1$ (όπου $\|R'\|_1$ η L_1 νόρμα του R'). Διαφορετικά: $\mathbf{R}' = c(\mathbf{G}\mathbf{R}' + \mathbf{E})$. Εφόσον $\|\mathbf{R}'\|_1 = 1$, ο τύπος μπορεί να γραφτεί και ως εξής: $\mathbf{R}' = c(\mathbf{G} + \mathbf{E} \times \mathbf{1})\mathbf{R}'$ όπου $\mathbf{1}$ διάνυσμα με όλα τα στοιχεία του ίσα με τη μονάδα. Επομένως το διάνυσμα $\mathbf{R}'$ αποτελεί ιδιοδιάνυσμα της μήτρας $\mathbf{G} + \mathbf{E} \times \mathbf{1}$.

Το μοντέλο του τυχαίου περιηγητή

Υιοθετείται το μοντέλο του «τυχαίου περιηγητή» (random surfer). Ο τυχαίος περιηγητής αποτελεί ένα ιδεατό ον που ξεκινάει την περιήγηση στον ιστό από μια τυχαία ιστοσελίδα. Στη συνέχεια ακολουθεί τυχαία κάποιο σύνδεσμο από την εκάστοτε ιστοσελίδα που επισκέπτεται. Η διαδικασία επαναλαμβάνεται συνεχώς, μέχρι ο τυχαίος περιηγητής να κουραστεί. Τότε αρχίζει μια νέα περιήγηση απο κάποια τυχαία σελίδα του Διαδικτύου, συμπεριφερόμενος κατά τον ίδιο τρόπο, μέχρι και πάλι να κουραστεί. Η διαδικασία αυτή επαναλαμβάνεται επ' άπειρον. Η τιμή PageRank αντανακλά την πιθανότητα με την οποία ο τυχαίος περιηγητής, στη διαδικασία περιήγησης, βρίσκεται στη δεδομένη ιστοσελίδα.

Ένας πραγματικός περιηγητής, όταν εισαχθεί σε έναν ατέρμονα βρόχο σελίδων, είναι απίθανο να συνεχίσει να κάνει κλικ στους συνδέσμους, οπότε θα επιλέξει να επισκεφθεί μία εκ νέου τυχαία αφετηρία για την επανέναρξη της πλοήγησής του στο Διαδίκτυο. Ο παράγοντας E, όπως ορίσθηκε παραπάνω, μπορεί να θεωρηθεί ως ένας τρόπος μοντελοποίησης αυτής της συμπεριφοράς, κατά την οποία ο περιηγητής περιοδικά «χάνει το ενδιαφέρον του» και παύει να κάνει κλικ διαδοχικά στους συνδέσμους, επιλέγοντας νέα

αφετηρία, η οποία επιλέγεται βάσει της κατανομής του E. Έτσι, ο εν λόγω παράγοντας θεωρείται μία παράμετρος του μοντέλου, ενώ διαφορετικές τιμές του E μπορούν να έχουν ως αποτέλεσμα εξατομικευμένα συστήματα PageRank.

Υπολογισμός του PageRank

Στη συγκεκριμένη ενότητα παρουσιάζεται ο τύπος του επαναληπτικού υπολογισμού του αλγόριθμου PageRank καθώς και ο τρόπος με τον οποίο ο τύπος αυτός αντιμετωπίζει το πρόβλημα των ιστοσελίδων χωρίς εξερχόμενους συνδέσμους και υλοποιεί το μοντέλο του τυχαίου περιηγητή.

Ιστοσελίδες χωρίς εξερχόμενους συνδέσμους (dangling nodes)

Ένας εναλλακτικός τρόπος να αντιμετωπίσουμε το πρόβλημα των ιστοσελίδων χωρίς εξερχόμενους συνδέσμους, οι οποίες δε μεταφέρουν το σκορ που λαμβάνουν σε άλλες, είναι να θεωρήσουμε ότι οι ιστοσελίδες αυτές έχουν εξερχόμενους συνδέσμους προς όλες τις υπόλοιπες ιστοσελίδες. Αν επανέλθουμε στο μοντέλο του τυχαίου περιηγητή, πραγματοποιώντας την αλλαγή που αναφέραμε, όταν αυτός βρεθεί σε μία τέτοια ιστοσελίδα ξεκινά την πλοήγησή του από την αρχή. Αυτό επιτυγχάνεται αν στην τετραγωνική μήτρα $\mathbf{G}$ του διαδικτυακού γράφου αντικαταστήσουμε όλες τις στήλες με μηδενικά, με την ποσότητα $1/\#$ιστοσελίδων του διαδικτυακού γράφου. Με τον τρόπο αυτό, απαλείφουμε από τον τύπο υπολογισμού του αλγόριθμου Pagerank τον παράγοντα κανονικοποίησης c.

Συντελεστής απόσβεσης (damping factor)

Το σκορ Pagerank μιας σελίδας είναι η πιθανότητα ο τυχαίος χρήστης να επισκεφθεί τη σελίδα αυτή. Ο πρώτος και ο δεύτερος όρος του τύπου 2.3 αντιστοιχούν στα ενδεχόμενα ο χρήστης να συνεχίσει την αναζήτησή του ανοίγοντας τους συνδέσμους της τρέχουσας ιστοσελίδας ή να ξεκινήσει από την αρχή αντίστοιχα, τα οποία είναι αμοιβαίως αποκλειόμενα. Ενσωματώνουμε λοιπόν αυτή τη λογική στον τύπο χρησιμοποιώντας έναν συντελεστή απόσβεσης a. Επομένως, αν η πρώτη πιθανότητα είναι a, η δεύτερη είναι $1 - a$. Η συνηθέστερη τιμή που παίρνει ο συντελεστής απόσβεσης είναι 0.85, η οποία αντανακλά τη θεώρηση ότι όταν κάνουμε περιήγηση στο διαδίκτυο, ακολουθούμε κατά μέσο όρο 6 συνδέσμους πριν αρχίσουμε εκ νέου μια τυχαία περιήγηση.

Με βάση τα παραπάνω, ο τύπος του επαναληπτικού αλγόριθμου διαμορφώνεται ως εξής:

$$\mathbf{p_{i+1}} = a(\mathbf{G} + \frac{\mathbf{1} \cdot \mathbf{d}^T}{N})\mathbf{p_i} + (1 - a)\frac{\mathbf{1}}{N} \tag{2.4}$$

όπου $\mathbf{1}$ διάνυσμα με όλα τα στοιχεία του ίσα με τη μονάδα, $\mathbf{d}$ διάνυσμα με τιμή ίση με 1 αν ο αντίστοιχος κόμβος δεν έχει εξερχόμενες ακμές και 0 αν έχει και a συντελεστής απόσβεσης. Οι γραμμές και οι στήλες του πίνακα $\mathbf{1} \cdot \mathbf{d}^T$ αντιπροσωπεύουν ιστοσελίδες, με τα στοιχεία των στηλών που αντιστοιχούν σε ιστοσελίδες χωρίς συνδέσμους, να είναι ίσα με τη μονάδα, ενώ αυτά των υπόλοιπων στηλών, ίσα με το μηδέν.

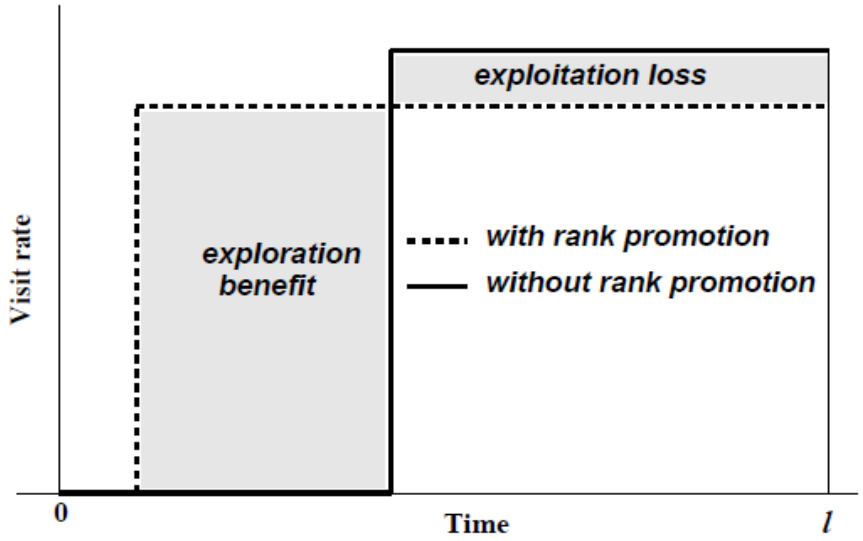
Οι επαναλήψεις πραγματοποιούνται μέχρι το σφάλμα της μετρικής να μειωθεί κάτω από μια ορισμένη τιμή. Το τελικό διάνυσμα $\mathbf{p}$ είναι κανονικοποιημένο ώστε το άθροισμα των στοιχείων του να ισούται με τη μονάδα.

2.1.1.2 Αλγόριθμος κατάταξης με είσοδο τυχαία κατάταξη των ιστοσελίδων

Χρησιμοποιώντας τον αλγόριθμο PageRank σαν μέθοδο κατάταξης ιστοσελίδων, κάνουμε την υπόθεση ότι η δημοτικότητα μιας ιστοσελίδας είναι στενά συσχετισμένη με την ποιότητά της. Σαν ποιότητα μιας σελίδας ορίζεται το ενδιαφέρον που θα έδειχναν οι χρήστες για τη σελίδα αυτή αν γνώριζαν την ύπαρξή της, ενώ σαν ενδιαφέρον ορίζεται η δημιουργία υπερσυνδέσμων προς αυτήν από άλλες ιστοσελίδες. Παρ' όλα αυτά, η αλληλοσυσχέτιση μεταξύ της ποιότητας και της δημοτικότητας των πρόσφατα δημιουργημένων ιστοσελίδων είναι αρκετά ασθενής, επειδή δεν έχουν δεχτεί ακόμα μεγάλο πλήθος επισκέψεων και επομένως εισερχόμενων συνδέσμων. Επιπλέον, εφόσον οι ιστοσελίδες που επισκέπτεται ένας χρήστης καθορίζονται από τα πρώτα κιόλας αποτελέσματα της μηχανής αναζήτησης που χρησιμοποιεί, νέες ιστοσελίδες που είναι σημαντικές καθυστερούν ακόμα περισσότερο να γίνουν δημοφιλείς.

Για τους παραπάνω λόγους, στην εργασία [28] προτείνεται η τελική κατάταξη των ιστοσελίδων σε μια μηχανή αναζήτησης να μην γίνεται εξ' ολοκλήρου με βάση κάποιον καθορισμένο αλγόριθμο αλλά, ως έναν βαθμό, με τυχαίο τρόπο. Με αυτόν τον τρόπο, δίνεται η ευκαιρία στις νέες ιστοσελίδες να γίνουν πιο γρήγορα δημοφιλείς. Αν όμως το ποσοστό με το οποίο η κατάταξη πραγματοποιείται με τυχαίο τρόπο αυξηθεί κατά πολύ, τότε αυτή δεν θα ανταποκρίνεται καθόλου στη δημοτικότητα των ιστοσελίδων. Δημιουργείται επομένως το δίλημμα μεταξύ της προώθησης των νέων ιστοσελίδων που πραγματοποιείται με την τυχαία κατάταξη (η εργασία αναφέρεται στη διαδικασία αυτή σαν *exploration*) και της υποβάθμισης των ήδη ποιοτικών παλαιότερα δημοσιευμένων ιστοσελίδων (η εργασία αναφέρεται στη διαδικασία αυτή σαν *exploitation*). Πρέπει λοιπόν να απαντηθούν τα παρακάτω ερωτήματα:

- *Ποιες σελίδες πρέπει να μπουν στο σύνολο αυτών που πρόκειται να προωθηθούν;*
- *Ποιες σελίδες πρέπει μην υποβαθμιστούν μετά την προώθηση του προηγούμενου συνόλου;*
- *Ποιος πρέπει να είναι τελικά ο λόγος μεταξύ της προώθησης και της υποβάθμισης των ιστοσελίδων;*



Σχήμα 2.5: Η προώθηση των νέων ιστοσελίδων (exploration) έναντι της υποβάθμισης των παλιών (exploitation).

Το Σχήμα 2.5 δείχνει την καμπύλη της δημοτικότητας μιας σημαντικής ιστοσελίδας σε σχέση με τον χρόνο. Παρατηρούμε ότι όταν χρησιμοποιούμε μηχανισμό προώθησης νέων ιστοσελίδων, η ιστοσελίδα αυτή γίνεται πιο γρήγορα δημοφιλής μιας και είναι πιο ψηλά στην κατάταξη και εντοπίζεται γρηγορότερα από τους χρήστες του Διαδικτύου. Παρ' όλα αυτά, μετά το πέρασμα κάποιου χρονικού διαστήματος, η δημοτικότητά της είναι χαμηλότερη από αυτή που θα μπορούσε να έχει, επειδή με τον ίδιο τρόπο προωθούνται και άλλες νέες ιστοσελίδες.

Δημιουργείται επομένως η ανάγκη εξισορρόπησης των αποτελεσμάτων μεταξύ της προώθησης των νέων ιστοσελίδων και υποβάθμισης των ήδη δημοφιλών, με σκοπό να δημιουργήσουμε τα πιο αξιόπιστα αποτελέσματα. Ορίζουμε σε αυτό το σημείο τις παρακάτω ποσότητες:

Ορισμός 2.1. *TBP (Time to become popular) = Ο χρόνος που απαιτείται ώστε μια νέα, σημαντική ιστοσελίδα να γίνει δημοφιλής, να αποκτήσει δηλαδή μεγάλο σκορ κατάταξης.*

Όταν η τιμή της ποσότητας αυτής ελαχιστοποιείται τότε έχει επιτευχθεί μία αποτελεσματική κατάταξη, επειδή οι νέες ιστοσελίδες προωθήθηκαν τόσο, ώστε να εντοπιστούν εγκαίρως από τους χρήστες και να γίνουν γρήγορα δημοφιλείς.

Ορισμός 2.2. *QPC = (Quality per click) Η μέση ποιότητα των ιστοσελίδων που επισκέπτονται οι χρήστες μετά το πέρασμα μεγάλης χρονικής περιόδου.*

$$QPC = \lim_{t\to\infty} \frac{\sum_{t_l=0}^{t} \sum_{p\in\mathcal{P}} (V_u(p,t_l) \cdot Q(p))}{\sum_{t_l=0}^{t} (\sum_{p\in\mathcal{P}} V_u(p,t_l))} \tag{2.5}$$

όπου $V_u(p,t_l)$ ο αριθμός των επισκέψεων της ιστοσελίδας p το χρονικό διάστημα t_l και $Q(p)$ η ποιότητα της ιστοσελίδας p.

Όταν η τιμή της ποσότητας αυτής μεγιστοποιείται, έχει επιτευχθεί μία αποτελεσματική κατάταξη. Αυτό συμβαίνει γιατί έχουμε καταφέρει να τοποθετήσουμε ψηλότερα στην κατάτα-

ξη τις πιο σημαντικές ιστοσελίδες με αποτέλεσμα όποια ιστοσελίδα και να επισκέπτονται οι χρήστες να είναι σημαντική.

Περιγραφή αλγόριθμου

Έστω το σύνολο όλων των ιστοσελίδων $\mathcal{P}$ και το υποσύνολό του $\mathcal{P}_p \subseteq \mathcal{P}$ το οποίο περιέχει τις ιστοσελίδες που έχουν επιλεχτεί να προωθηθούν σύμφωνα με έναν προκαθορισμένο κανόνα. Οι ιστοσελίδες του $\mathcal{P}_p$ κατατάσσονται τυχαία δημιουργώντας τη λίστα $\mathcal{L}_p$. Οι υπόλοιπες ιστοσελίδες $(\mathcal{P} - \mathcal{P}_p)$ κατατάσσονται με ντετερμινιστικό τρόπο (αλγόριθμος PageRank) και δημιουργούν τη λίστα $\mathcal{L}_d$. Οι δύο λίστες που δημιουργήθηκαν συγχωνεύονται έτσι ώστε να δημιουργήσουν την τελική λίστα $\mathcal{L}$ σύμφωνα με την ακόλουθη διαδικασία:

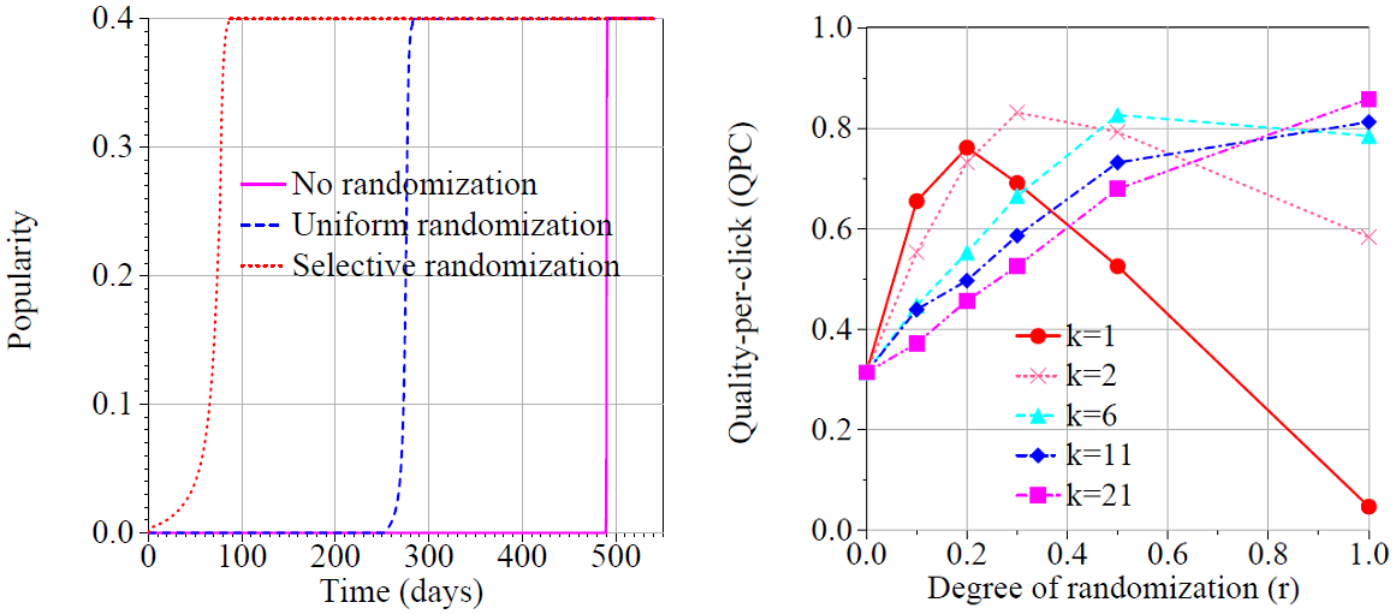
1. Τα πρώτα $k-1$ στοιχεία της λίστας $\mathcal{L}_d$ εισάγονται στην αρχή της λίστας $\mathcal{L}$ διατηρώντας την κατάταξή τους.

2. Για κάθε μία από τις υπόλοιπες θέσεις της λίστας $\mathcal{L}$ $(i = k, k+1, ..., n)$ η ιστοσελίδα που πρόκειται να εισαχθεί καθορίζεται ρίχνοντας ένα μεροληπτικό νόμισμα με πιθανότητα r το επόμενο στοιχείο να προέρχεται από την αρχή της λίστας $\mathcal{L}_p$ και $1-r$ το επόμενο στοιχείο να προέρχεται από την αρχή της λίστας $\mathcal{L}_d$ μέχρι κάποια από τις δύο να αδειάσει και στο τέλος να εισαχθούν τα στοιχεία της μη κενής λίστας.

Στη διαδικασία που περιγράφηκε αναφέρθηκαν οι παρακάτω παράγοντες:

- Το σύνολο $\mathcal{P}_p$ των ιστοσελίδων που πρόκειται να προωθηθούν το οποίο μπορεί να διαμορφωθεί με δύο διαφορετικές πολιτικές προώθησης:
  1. Την «ομοιογενή», κατά την οποία όλες οι ιστοσελίδες μπορεί να ανήκουν στο σύνολο $\mathcal{P}_p$ με πιθανότητα r.
  2. Την «επιλεκτική», κατά την οποία στο σύνολο $\mathcal{P}_p$ ανήκουν μόνο οι ιστοσελίδες που δεν είναι ακόμα γνωστές στους χρήστες (έχουν πλήθος επισκέψεων μικρότερο του v και ηλικία μικρότερη του l).
- Η θέση k μετά από την οποία πρόκειται να εισάγονται οι ιστοσελίδες του συνόλου $\mathcal{P}_p$.
- Το ποσοστό r σύμφωνα με το οποίο η τελική λίστα διαμορφώνεται από ιστοσελίδες που ανήκουν στο σύνολο $\mathcal{P}_p$ έναντι του συνόλου $\mathcal{P} - \mathcal{P}_p$,.

Οι παράγοντες αυτοί λαμβάνουν τέτοιες τιμές ώστε οι ποσότητες TBC και QBC να μεγιστοποιούνται.



Σχήμα 2.6: Ο χρόνος που απαιτείται για να γίνει γνωστή μια ιστοσελίδα μέσης ποιότητας ανάλογα με τον τρόπο προώθησης των «μη γνωστών» ιστοσελίδων: χωρίς, με ομοιόμορφη ή επιλεκτική προώθηση και η ποσότητα QBC για διάφορες τιμές του k,r.

Σύμφωνα με τα διαγράμματα του σχήματος 2.1.1.2, τα οποία προέρχονται από την εργασία [28], η ποσότητα TBP μεγιστοποιείται όταν χρησιμοποιούμε πολιτική επιλεκτικής προώθησης ενώ η ποσότητα QBC για $k = 2$ και $r = 0.3$.

Παράδειγμα

Έστω λοιπόν ότι το σύνολο των ιστοσελίδων της εξεταζόμενης μηχανής αναζήτησης είναι 100, εκ των οποίων οι 9 έχουν δημιουργηθεί τα τελευταία l χρόνια και έχουν λιγότερες από v επισκέψεις. Τα βήματα του προτεινόμενου αλγόριθμου είναι τα εξής:

1. Αποτιμούμε τις παραμέτρους r, k με τις τιμές για τις οποίες έχουμε τα καλύτερα αποτελέσματα ($r = 0, 3, k = 2$).
2. Διαμορφώνουμε το σύνολο $\mathcal{P}_p$ ανάλογα με την πολιτική προώθησης:
   - Ομοιογενής: Από τις 100 ιστοσελίδες επιλέγονται τυχαία οι $100r = 30$.
   - Επιλεκτική: Το σύνολο $\mathcal{P}_p$ αποτελείται από τις 9 ιστοσελίδες που έχουν δημιουργηθεί τα τελευταία l χρόνια και έχουν λιγότερες από v επισκέψεις.
3. Ταξινομούμε το σύνολο $\mathcal{P} - \mathcal{P}_p$ το οποίο αποτελείται από 70 δημοσιεύσεις στην περίπτωση της ομοιόμορφης προώθησης και από 91 σε αυτήν της επιλεκτικής, σύμφωνα με κάποιον αλγόριθμο κατάταξης (π.χ. PageRank). Προκύπτει λοιπόν η λίστα $L_{\mathcal{P}-\mathcal{P}_p}$.
4. Στην πρώτη θέση της κατάταξης βάζουμε το πρώτο στοιχείο της λίστας $L_{\mathcal{P}-\mathcal{P}_p}$.
5. Στην περίπτωση της ομοιόμορφης προώθησης οι υπόλοιπες 99 θέσεις καταλαμβάνονται από το σύνολο $\mathcal{P}_p$ και την λίστα $L_{\mathcal{P}-\mathcal{P}_p}$ έτσι ώστε ανά 10 ιστοσελίδες

της τελικής κατάταξης, οι 7 να αποτελούνται από την ταξινομημένη λίστα $L_{\mathcal{P}-\mathcal{P}_p}$ με τη σειρά που έχουν σε αυτήν και οι 3 από το σύνολο $\mathcal{P}_p$. Στην περίπτωση της επιλεκτικής προώθησης οι επόμενες 30 θέσεις καταλαμβάνονται όπως και στην περίπτωση της ομοιόμορφης προώθησης ενώ οι τελευταίες 69 από τις υπόλοιπες ιστοσελίδες της λίστας $L_{\mathcal{P}-\mathcal{P}_p}$.

2.2 Κατάταξη των δημοσιεύσεων

Η εφαρμογή Diana MirPub αποτελεί μηχανή αναζήτησης επιστημονικών εργασιών. Κάθε εργασία μπορεί να αναφέρει άλλες, δημιουργώντας έτσι μια δομή γράφου. Στον γράφο αυτόν, οι κόμβοι αποτελούνται απο τις ίδιες τις δημοσιεύσεις, ενώ οι ακμές απο τις παραπομπές μιας δημοσίευσης προς μία άλλη. Ο γράφος μπορεί να αντιστοιχηθεί με αυτόν των ιστοσελίδων, αφού μια παραπομπή υποδηλώνει το ενδιαφέρον των συγγραφέων για μία άλλη δημοσίευση, όπως ένας σύνδεσμος μιας ιστοσελίδας υποδηλώνει το ενδιαφέρον, ή τη σχέση των διαχειριστών της με μία άλλη.

Η κύρια διαφορά του διαδικτυακού γράφου με αυτόν των επιστημονικών εργασιών είναι ότι οι εξερχόμενες ακμές ενός κόμβου μπορούν να μεταβάλλονται εφόσον το αρχικό περιεχόμενο μιας ιστοσελίδας, άρα και οι σύνδεσμοι που διαθέτει, μπορεί να αλλάξει. Αντίθετα, το περιεχόμενο των επιστημονικών εργασιών παραμένει αμετάβλητο μετά τη δημοσίευσή τους, μαζί με τη βιβλιογραφία τους. Αυτό έχει σαν αποτέλεσμα, οι επιστημονικές εργασίες να παραπέμπουν μόνο σε προγενέστερες από αυτές εργασίες.

Γι' αυτό το λόγο, εργασίες αντίστοιχης σημαντικότητας δεν είναι άμεσα συγκρίσιμες με βάση το πλήθος των παραπομπών προς αυτές (και την ποιότητα των παραπομπών αυτών) όταν το έτος δημοσίευσής τους διαφέρει κατά πολύ. Το φαινόμενο αυτό, μπορεί να συναντάται και στις δύο περιπτώσεις, αλλά όσον αφορά στις ιστοσελίδες, περιορίζεται στις πρόσφατα δημιουργημένες, για τις οποίες δεν έχουν προλάβει να δημιουργηθούν εισερχόμενοι σύνδεσμοι. Οι ιστοσελίδες, όταν αρχίσουν να γίνονται δημοφιλείς, θα αποκτήσουν εισερχόμενους συνδέσμους και από προγενέστερες σελίδες, κάτι το οποίο δεν συμβαίνει στην περίπτωση των εργασιών.

Είναι αναγκαίο λοιπόν, για την κατάταξη των δημοσιεύσεων, να υλοποιηθούν και να αξιολογηθούν παραλλαγές των ευρέως χρησιμοποιούμενων αλγόριθμων κατάταξης των ιστοσελίδων του Διαδικτύου, που να λαμβάνουν υπόψη τις διαφορές στη δομή των δύο γράφων.

2.3 Μετρικές κατάταξης περιοδικών

Η επιστημονική βιβλιογραφία αποτελεί ένα τεράστιο δίκτυο από εργασίες που συνδέονται μεταξύ τους με τις αναφορές στις βιβλιογραφίες των εργασιών αυτών. Η δομή αυτού του δικτύου αντανακλά τις αποφάσεις των επιστημόνων για το ποια έγγραφα είναι σχετικά και σημαντικά για τις εργασίες τους. Γι' αυτό το λόγο από τη δομή του συγκεκριμένου δικτύου μπορεί να εξαχθεί πληροφορία για τη δημοτικότητα και το κύρος των αντίστοιχων περιοδικών καθώς και για τη σχέση μεταξύ των διάφορων επιστημονικών κλάδων.

2.3.1 Impact factor

Το impact factor ενός επιστημονικού περιοδικού αποτελεί μία μετρική που αντικατοπτρίζει τον μέσο αριθμό των παραπομπών σε πρόσφατες δημοσιεύσεις του εν λόγω περιοδικού. Η μετρική αυτή υπολογίζεται ετησίως και χρησιμοποιείται για τη σύγκριση διαφορετικών περιοδικών.

Σε οποιοδήποτε δεδομένο έτος, ο παράγοντας επιρροής (impact factor) ενός περιοδικού είναι ο μέσος αριθμός των αναφορών που λαμβάνονται το έτος αυτό, προς τα άρθρα που δημοσιεύτηκαν στο περιοδικό, κατά τη διάρκεια των t προηγούμενων ετών.

$$\mathbf{IF}(v, y : t) = \frac{Cited(\cup_{i=1..t} V_{y-i}, y)}{|\cup_{i=1..t} V_{y-i}|} \tag{2.6}$$

όπου V_y το σύνολο των άρθρων που δημοσιεύτηκαν στο περιοδικό v το έτος y, t το χρονικό διάστημα μέσα στο οποίο ανήκει η ένωση των συνόλων V_y (συνήθως ισούται με 2).
Η συνάρτηση $Cited(A, y)$ μετράει τις αναφορές προς τα άρθρα του συνόλου A των άρθρων που δημοσιεύτηκαν το έτος y.

Επομένως, η μετρική αυτή για ένα περιοδικό το έτος y δεν μπορεί να είναι γνωστή πριν το έτος $y + 1$, έτσι ώστε να συγκεντρωθούν όλες οι αναφορές προς τις δημοσιεύσεις του που πραγματοποιήθηκαν το έτος y. Συνεπώς, νέα περιοδικά τα οποία δημοσιεύονται για πρώτη φορά, θα λάβουν τιμή της μετρικής τρία χρόνια μετά τη δημοσίευση της πρώτης έκδοσής τους.

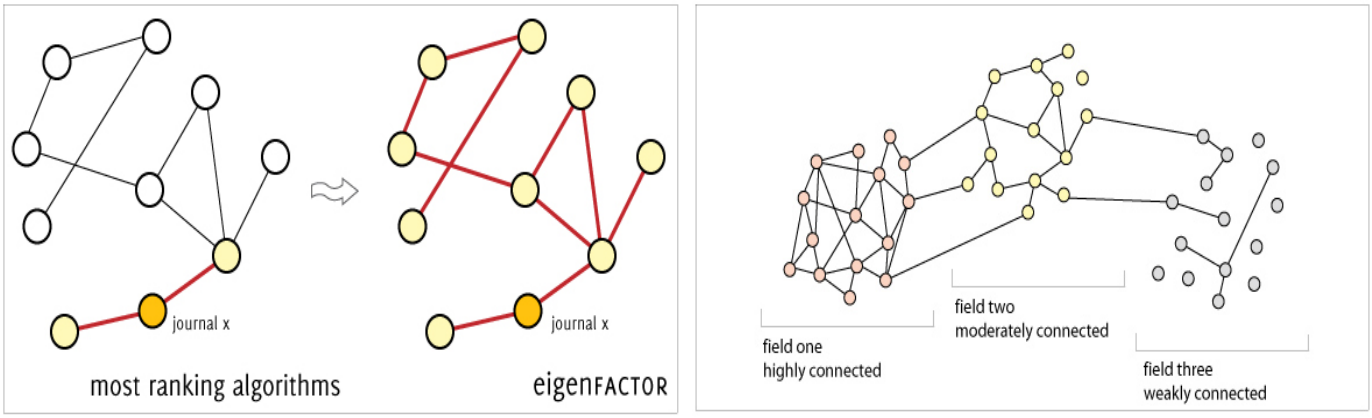
Η μετρική αυτή μπορεί να μην είναι αξιόπιστη όταν χρησιμοποιείται για τη σύγκριση περιοδικών διαφορετικού επιστημονικού περιεχομένου, μιας και το ποσοστό των παραπομπών που πραγματοποιούνται στους διαφορετικούς επιστημονικούς κλάδους μπορεί να διαφέρει σημαντικά. Θα ήταν λάθος επίσης να χρησιμοποιήσουμε μόνο το impact factor ενός περιοδικού για να αξιολογήσουμε κάθε άρθρο που είναι δημοσιευμένο σε αυτό. Για παράδειγμα, το 90% του αντίστοιχου σκορ για το περιοδικό Nature το έτος 2014 βασίστηκε στο 1\4 των άρθρων του. Αυτό γίνεται ακόμα πιο αισθητό από τη στιγμή που τα άρθρα των περιοδικών είναι ελεύθερα διαθέσιμα στο Διαδίκτυο, με αποτέλεσμα η ανάγνωση ενός άρθρου να συνδυάζεται όλο και λιγότερο με την ανάγνωση και των υπόλοιπων άρθων ενός περιοδικού.

2.3.2 Eigenfactor - Article Influence score

Ο αλγόριθμος Eigenfactor κατατάσσει τα περιοδικά με βάση τη κύρος τους όπως ο αλγόριθμος PageRank τις ιστοσελίδες, ενώ η μετρική Article Influence score πραγματοποιεί κανονικοποίηση των αποτελεσμάτων του Eigenfactor. Σύμφωνα με την προσέγγιση αυτή, τα περιοδικά θεωρούνται σημαντικά όταν αναφέρονται συχνά από άλλα σημαντικά περιοδικά. Οι αλγόριθμοι αυτοί εφαρμόζονται στον γράφο που έχει σαν κόμβους τα περιοδικά και ακμές τις παραπομπές των άρθρων που είναι δημοσιευμένα σε κάθε περιοδικό προς άρθρα άλλων περιοδικών. Σε αντίθεση με τους περισσότερους αλγόριθμους κατάταξης περιοδικών, το σκορ ενός περιοδικού υπολογίζεται με βάση τη δομή ολόκληρου του δικτύου και όχι μόνο με βάση τις άμεσες αναφορές σε άρθρα του.

Η διαφορά της μετρικής Eigenfactor με το Impact factor των περιοδικών έγκειται στο γεγονός ότι η πρώτη κατατάσσει τα περιοδικά με βάση το κύρος τους και όχι απλά τη δημοτικότητά τους. Δίνει έμφαση δηλαδή στους εισερχόμενους συνδέσμους των περιοδικών από σημαντικά άρθρα άλλων περιοδικών.

Το πλήθος και η συχνότητα των αναφορών που λαμβάνει ένα άρθρο διαφέρει ανάλογα με τον κλάδο της επιστήμης στον οποίο κατατάσσεται. Για παράδειγμα, ο μέσος όρος των παραπομπών προς ένα άρθρο σε ένα κορυφαίο περιοδικό βιολογίας μπορεί να λάβει 10-30 αναφορές μέσα σε δύο χρόνια. Αντίθετα, ο αντίστοιχος μέσος όρος κατά την ίδια περίοδο ενός άρθρου σχετικού με την επιστήμη των μαθηματικών είναι 2 αναφορές. Με τη χρήση λοιπόν του συνόλου του δικτύου παραπομπών, εξασφαλίζεται η καλύτερη σύγκριση μεταξύ περιοδικών διαφορετικού επιστημονικού αντικειμένου.



Σχήμα 2.7: Ο αλγόριθμος Eigenfactor χρησιμοποιεί το σύνολο των παραπομπών όλων των περιοδικών για την αποτίμηση του σκορ κατάταξης κάθε περιοδικού. Έτσι λαμβάνονται υπόψη οι διαφορές στην πυκνότητα των υπογράφων μεταξύ των επιστημονικών αντικειμένων.

Ο υπολογισμός των μετρικών Eigenfactor και Article Influence score διατίθεται στην ιστοσελίδα http://www.eigenfactor.org[5]. Στη συγκεκριμένη ιστοσελίδα, για τον υπολογισμό των σκορ κάθε επιστημονικού περιοδικού ο γράφος εμπλουτίστηκε και με άρθρα εφημερίδων καθώς και διπλωματικές στον χώρο των φυσικών και κοινωνικών επιστημών έτσι ώστε να λαμβάνεται υπόψη το κύρος του κάθε περιοδικού με βάση και αυτές τις δημοσιεύσεις και όχι μόνο τα επιστημονικά άρθρα.

Σε πολλούς τομείς της έρευνας, δεν δημιουργούνται αναφορές προς τα αντίστοιχα επιστημονικά άρθρα παρά μόνο μετά από αρκετά χρόνια από τη δημοσίευσή τους. Ως εκ τούτου, τα συμπεράσματα για το κύρος ενός περιοδικού που λαμβάνουν υπόψη τις αναφορές κατά τα πρώτα δύο έτη μετά τη δημοσίευση μπορεί να είναι παραπλανητικά. Γι' αυτό το λόγο τα σκορ Eigenfactor και Article Influence υπολογίζονται με βάση τις αναφορές που έλαβε κάθε επιστημονικό άρθρο (συνεπώς και το αντίστοιχο περιοδικό) κατά τη διάρκεια μιας περιόδου πέντε ετών.

Το σκορ Eigenfactor ενός περιοδικού είναι μια εκτίμηση του ποσοστού του χρόνου που δαπανά ενάς τυχαίος αναγνώστης στο εν λόγω περιοδικό. Ο αλγόριθμος Eigenfactor αντιστοιχεί σε ένα απλό μοντέλο όπου οι αναγνώστες ακολουθούν τις αλυσίδες των παραπομπών,

καθώς κινούνται μεταξύ των περιοδικών. Έστω ότι ένας ερευνητής πηγαίνει στη βιβλιοθήκη και επιλέγει ένα άρθρο του περιοδικού τυχαία. Μετά την ανάγνωση του άρθρου, ο ερευνητής επιλέγει τυχαία μία από τις αναφορές του άρθρου. Στη συνέχεια προχωρά στο αντίστοιχο περιοδικό, διαβάζει ένα τυχαίο άρθρο εκεί, και επιλέγει μια παραπομπή κ.ο.κ. Ο ερευνητής ακολουθεί επ' άπειρον την συγκεκριμένη διαδικασία. Το χρονικό διάστημα που ο ερευνητής περνά σε κάθε περιοδικό μας δίνει ένα μέτρο της σημαντικότητας του εν λόγω περιοδικού εντός του δικτύου των ακαδημαϊκών αναφορών.

Τρόπος υπολογισμού Eigenfactor και Article Influence

Έχοντας σαν δεδομένο το σύνολο των επιστημονικών άρθρων, τις αναφορές τους, το περιοδικό και τη χρονολογία δημοσίευσής τους, δημιουργούμε τη μήτρα $\mathbf{Z}$ όπου:
$\mathbf{Z}_{ij}$ = Το σύνολο των παραπομπών των άρθρων που είναι δημοσιευμένα στο περιοδικό j το τρέχον έτος, προς τα άρθρα του περιοδικού i τα οποία είναι δημοσιευμένα σε διάστημα 5 ετών πριν το τρέχον έτος. Επομένως, η τετραγωνική μήτρα $\mathbf{Z}$ έχει διάσταση $n \times n$ όπου n το πλήθος των περιοδικών. Για παράδειγμα, ας θεωρήσουμε ότι έχουμε τα περιοδικά A, B, C, D, E και F, στα οποία είναι δημοσιευμένες συνολικά 14 εργασίες και οι μεταξύ τους παραπομπές δημιουργούν την παρακάτω μήτρα $\mathbf{Z}$ για τα περιοδικά τους:

	A	B	C	D	E	F
A	1	0	2	0	4	3
B	3	0	1	1	0	0
C	2	0	4	0	1	0
D	0	0	1	0	0	1
E	8	0	3	0	5	2
F	0	0	0	0	0	0

Επομένως, το σύνολο των παραπομπών μεταξύ των άρθρων του περιοδικού A είναι 1, οι παραπομπές των άρθρων του περιοδικού A προς αυτά του περιοδικού B και D είναι 0, προς τα άρθρα του περιοδικού C, 2, προς τα άρθρα του περιοδικού E, 4 και προς τα άρθρα του περιοδικού F, 3. Ομοίως και για τα υπόλοιπα περιοδικά.

Ορίζουμε τη μήτρα $\mathbf{H}$ η οποία αποτελεί μία τροποποίηση της $\mathbf{Z}$ αν αγνοήσουμε τις παραπομπές μεταξύ των άρθρων δημοσιευμένων στο ίδιο περιοδικό και κανονικοποιήσουμε κάθε στήλη με το άθροισμα των στοιχείων της. Άρα για το παραπάνω παράδειγμα η μήτρα $\mathbf{H}$ είναι η εξής:

	A	B	C	D	E	F
A	0	0	2/7	0	4/5	3/6
B	3/13	0	1/7	1	0	0
C	2/13	0	0	0	1/5	0
D	0	0	1/7	0	0	1/6
E	8/13	0	3/7	0	0	2/6
F	0	0	0	0	0	0

Ορίζουμε επίσης το διάνυσμα $\mathbf{a}$, όπου $\mathbf{a}_i$ ο αριθμός των άρθρων που δημοσιεύτηκαν στο περιοδικό i μέσα σε διάστημα 5 ετών πριν το τρέχον έτος προς το σύνολο των άρθρων όλων των περιοδικών, δημοσιευμένα το ίδιο διάστημα. Επομένως το διάνυσμα $\mathbf{a}$ είναι κανονικοποιημένο διάνυσμα που κάθε στοιχείο του υποδηλώνει το ποσοστό των άρθρων που είναι δημοσιευμένα στο αντίστοιχο περιοδικό. Για το παραπάνω παράδειγμα, δεδομένου ότι το σύνολο των δημοσιεύσεων ισούται με 14, το διάνυσμα $\mathbf{a}$ παίρνει την εξής μορφή:

	$\mathbf{a}_i$
A	3/14
B	2/14
C	5/14
D	1/14
E	2/14
F	1/14

Κάποια από τα περιοδικά της μήτρας $\mathbf{H}$ δεν έχουν εξερχόμενες ακμές, τα άρθρα τους δηλαδή δεν έχουν παραπομπές προς άρθρα άλλων περιοδικών. Κάθε στήλη της μήτρας $\mathbf{H}$ της οποίας τα στοιχεία είναι όλα μηδενικά αντιστοιχεί σε τέτοιο περιοδικό. Αντικαθιστούμε όλες αυτές τις στήλες της μήτρας $\mathbf{H}$ με το διάνυσμα $\mathbf{a}$ που υπολογίσαμε προηγουμένως και δημιουργούμε την τροποποιημένη μήτρα $\mathbf{H}'$.

Στο παράδειγμά μας η μήτρα $\mathbf{H}'$ είναι η παρακάτω:

	A	B	C	D	E	F
A	0	3/14	2/7	0	4/5	3/6
B	3/13	2/14	1/7	1	0	0
C	2/13	5/14	0	0	1/5	0
D	0	1/14	1/7	0	0	1/6
E	8/13	2/14	3/7	0	0	2/6
F	0	1/14	0	0	0	0

Ακολουθώντας την προσέγγιση του PageRank δημιουργούμε τη μήτρα $\mathbf{P}$ ως εξής:

$$\mathbf{P} = a\mathbf{H}' + (1-a)\mathbf{a}.e^T \tag{2.7}$$

όπου e^T διάνυσμα με όλα τα στοιχεία του ίσα με τη μονάδα και $\mathbf{a}.e^T$ μήτρα με στήλες ίσες με το διάνυσμα $\mathbf{a}$.

Επειδή η μήτρα $\mathbf{P}$ αποτελεί μία απεριοδική και αμείωτη Μαρκοβιανή αλυσίδα, σύμφωνα με το θεώρημα Perron - Frobenius, θα έχει ένα μοναδικό κύριο ιδιοδιάνυσμα π^* το οποίο αποτελεί και το μέτρο σύγκρισης του κύρους των περιοδικών. Θα μπορούσαμε να υπολογίσουμε το κύριο αυτό ιδιοδιάνυσμα κατευθείαν στη μήτρα $\mathbf{P}$, αλλά αυτό θα οδηγούσε σε σύνθετες πράξεις μιας και η μήτρα αυτή αποτελεί αρκετά πυκνό πίνακα. Εναλλακτικά, θα χρησιμοποιήσουμε μια προσέγγιση που πραγματοποιεί πράξεις στον αραιό πίνακα $\mathbf{H}$ εκτελώντας τον παρακάτω επαναληπτικό αλγόριθμο:

$$\pi^{(k+1)} = a\mathbf{H}\pi^k + [a\mathbf{d}.\pi^k + (1-a)]\mathbf{a} \tag{2.8}$$

όπου $\mathbf{d}$ το διάνυσμα που υποδηλώνει αν κάθε περιοδικό έχει (0) ή δεν έχει (1) εξερχόμενες ακμές.

Στο παράδειγμά μας το διάνυσμα $\mathbf{d}$ ορίζεται ως εξής:

	A	B	C	D	E	F
$\mathbf{d}_i$	0	1	0	0	0	0

Μετά από ένα πλήθος επαναλήψεων το διάνυσμα $\pi^{(k+1)}$ συγκλίνει στο ιδιοδιάνυσμα της μήτρας $\mathbf{P}$. Συνήθως απαιτούνται το πολύ 100 επαναλήψεις ώστε $\pi^{(k+1)} \approx \pi^*$ με $\pi^{(k+1)} - \pi^k < 0.00001$. Τα αποτελέσματα του αλγόριθμου είναι κανονικοποιημένα ώστε το συνολικό άθροισμα των σκορ να ισούται με 1.

Παρακάτω φαίνονται τα αποτελέσματα του αλγόριθμου για το παράδειγμα που χρησιμοποιήσαμε, ο οποίος συγκλίνει μετά από 16 επαναλήψεις, με είσοδο του αλγόριθμου το κανονικοποιημένο διάνυσμα $\pi_i^k = 1/n$ (όπου $n = 6$ το πλήθος των περιοδικών).

	π_i^*
A	0.3040
B	0.1636
C	0.1898
D	0.0466
E	0.2753
F	0.0206

Τέλος, οι μετρικές Eigenfactor $\mathbf{EF}_i$ και Article Influence $\mathbf{AI}_i$ για κάθε περιοδικό i ορίζονται ως εξής:

$$\mathbf{EF}_i = 100\frac{\mathbf{H}\pi^*}{\sum_j [\mathbf{H}\pi^*]_j} \tag{2.9}$$

$$\mathbf{AI}_i = 0.01\frac{\mathbf{EF}_i}{\mathbf{a}_i} \tag{2.10}$$

2.4 Σχετικές εργασίες

Σκοπός μας είναι να κατατάξουμε επιστημονικές εργασίες και όχι ιστοσελίδες. Επομένως ακόμα κι αν έχουμε σαν βάση τον αλγόριθμο Pagerank, πρέπει να λάβουμε υπόψη τις διαφορές μεταξύ του διαδικτυακού γράφου και αυτού των παραπομπών μεταξύ των επιστημονικών εργασιών. Παρακάτω, παρουσιάζονται παραλλαγές του αλγόριθμου PageRank που εφαρμόζονται για την κατάταξη εργασιών, τις οποίες χρησιμοποιήσαμε.

2.4.1 Μηχανισμός κατάταξης με βάση την προώθηση νέων δημοσιεύσεων και το impact factor του αντίστοιχου περιοδικού

Στην Εργασία [29] προτείνεται ένας αλγόριθμος κατάταξης επιστημονικών δημοσιεύσεων, η υλοποίηση του οποίου πρέπει να ικανοποιεί τους παρακάτω στόχους:

1. *Εργασίες για τις οποίες υπάρχουν παραπομπές στις βιβλιογραφίες σημαντικών εργασιών είναι σημαντικές.*

2. *Εργασίες δημοσιευμένες σε σημαντικά περιοδικά και συνέδρια είναι σημαντικές.*

3. *Δεν μπορούμε να αξιολογήσουμε τις πρόσφατα δημοσιευμένες εργασίες μόνο από το πλήθος των παραπομπών προς αυτές, από τη στιγμή που υπάρχει περίπτωση να λάβουν μεγάλο πλήθος παραπομπών στο μέλλον.*

Για να επιτευχθεί ο πρώτος στόχος χρησιμοποείται ο παρακάτω επαναληπτικός αλγόριθμός:

$$\mathbf{p_{i+1}} = a(\mathbf{G} + \mathbf{w} \cdot \mathbf{d}^T)\mathbf{p_i} + (1-a)\mathbf{w} \tag{2.11}$$

όπου $\mathbf{G}$ τετραγωνική μήτρα με τις γραμμές u και τις στήλες v να αντιστοιχούν στις δημοσιεύσεις και $\mathbf{G}_{u,v} = 1/N_v$, με N_v το σύνολο των παραπομπών της δημοσίευσης v, όταν υπάρχει παραπομπή από το v στο u και $\mathbf{G}_{u,v} = 0$ όταν δεν υπάρχει, $\mathbf{w}$ κανονικοποιημένο διάνυσμα με τιμή για όλες τις δημοσιεύσεις ίση με 1/#δημοσιεύσεων, $\mathbf{d}$ διάνυσμα με τιμή ίση με 1 αν ο αντίστοιχος κόμβος δεν έχει εξερχόμενες ακμές, διαφορετικά 0 και a συντελεστής απόσβεσης.

Ο τύπος 2.11 ουσιαστικά αποτελεί τον αλγόριθμο PageRank.

Για να επιτευχθεί ο δεύτερος στόχος, το διάνυσμα $\mathbf{w}$ διαμορφώνεται ώς εξής:

$$\mathbf{w}[i] = \frac{\mathbf{IF}(v_i, y_i : 5)}{\sum_{j=1}^{N} \mathbf{IF}(v_j, y_j : 5)} \tag{2.12}$$

όπου v_i το περιοδικό στο οποίο είναι δημοσιευμένο το άρθρο i, y_i το έτος δημοσίευσης του άρθρου i και $\mathbf{IF}(v_i, y : t)$ ο τύπος (2.6).

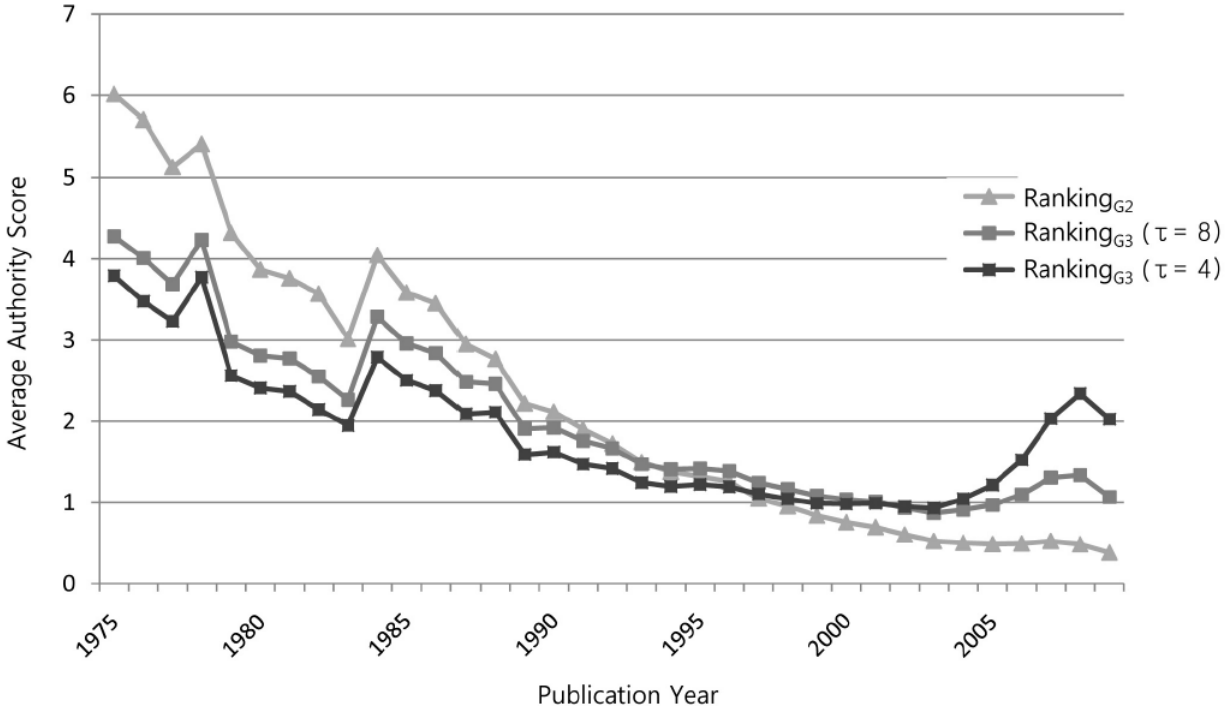
Για να επιτευχθεί ο τρίτος στόχος, πρέπει με κάποιο τρόπο να προωθηθούν οι νέες δημοσιεύσεις έναντι των παλιών στην κατάταξη που δημιουργείται. Γι' αυτό το λόγο χρησιμοποιείται ένας συντελεστής απόσβεσης ο οποίος για κάθε δημοσίευση p αποτιμάται ως εξής:

$$\rho_p = e^{-age(p)/\tau}/\tau \tag{2.13}$$

όπου τ παράγοντας γήρανσης (τ=4 ή τ=8) και $age(p)$ η ηλικία του p

και το τελικό διάνυσμα w διαμορφώνεται ως εξής:

$$\mathbf{w}[i] = \frac{\mathbf{IF}(v_i, y_i : 5) * \rho_{p_i}}{\sum_{j=1}^{N} \mathbf{IF}(v_j, y_j : 5) * \rho_{p_j}} \tag{2.14}$$



Σχήμα 2.8: Η επίδραση του παράγοντα γήρανσης στο τελικό σκορ των δημοσιεύσεων ανάλογα με τη χρονολογία δημοσίευσης. Το διάγραμμα προέρχεται από την εργασία [29], όπου τα δεδομένα που χρησιμοποιήθηκαν έχουν αντληθεί από την υπηρεσία ελεύθερων δεδομένων dblp [3], τον Μάρτιο του 2009.

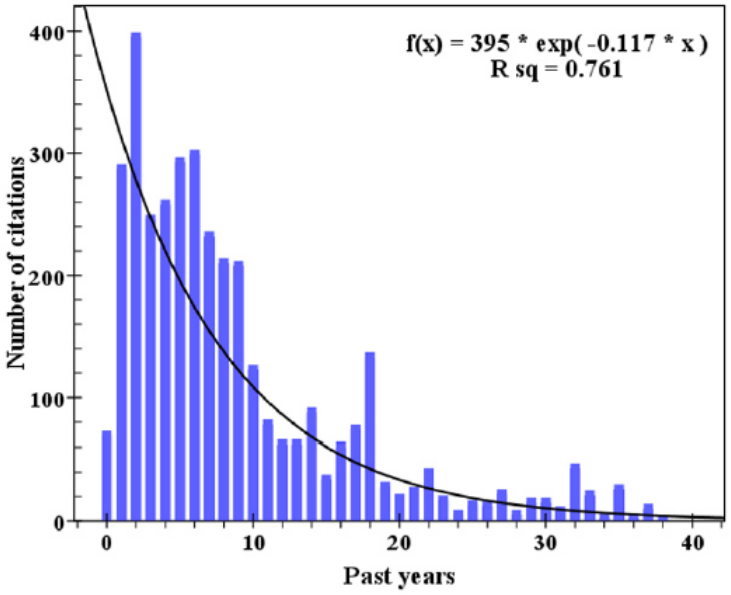
Η μέθοδος ουσιαστικά τροποποιεί το μοντέλο του τυχαίου περιηγητή στην περίπτωση που ξεκινάει μια νέα αναζήτηση ή φτάνει σε κόμβο χωρίς εξερχόμενες ακμές, δίνοντας περισσότερες πιθανότητες να επιλέξει πρόσφατες εργασίες, δημοσιευμένες σε δημοφιλή περιοδικά.

2.4.2 Αλγόριθμος κατάταξης σε γράφο με σταθμισμένες ακμές

Στην Εργασία [7] χρησιμοποιείται ο γράφος των επιστημονικών εργασιών που δημιουργείται από τις δημοσιεύσεις και τις παραπομπές τους. Σε κάθε ακμή του γράφου αυτού όμως, αντιστοιχίζεται ένα βάρος που προκύπτει από το συνδυασμό των εξής δύο παραγόντων: (α) το κύρος του περιοδικού στο οποίο δημοσιεύεται η εργασία που κάνει την αναφορά και (β) το μέγεθος του χρονικού διαστήματος που μεσολάβησε μεταξύ δημοσίευσης και δημιουργίας της αναφοράς.

Η λειτουργία του προτεινόμενου αλγόριθμου βασίζεται στις εξής παραδοχές:

- *Παραπομπές που προέρχονται από εργασίες δημοσιευμένες σε έγκυρα περιοδικά δείχνουν σε σημαντικές εργασίες*
- *Παραπομπές που δημιουργούνται σε μικρό χρονικό διάστημα μετά τη δημοσίευση της αντίστοιχης εργασίας δείχνουν σε σημαντικές εργασίες.*



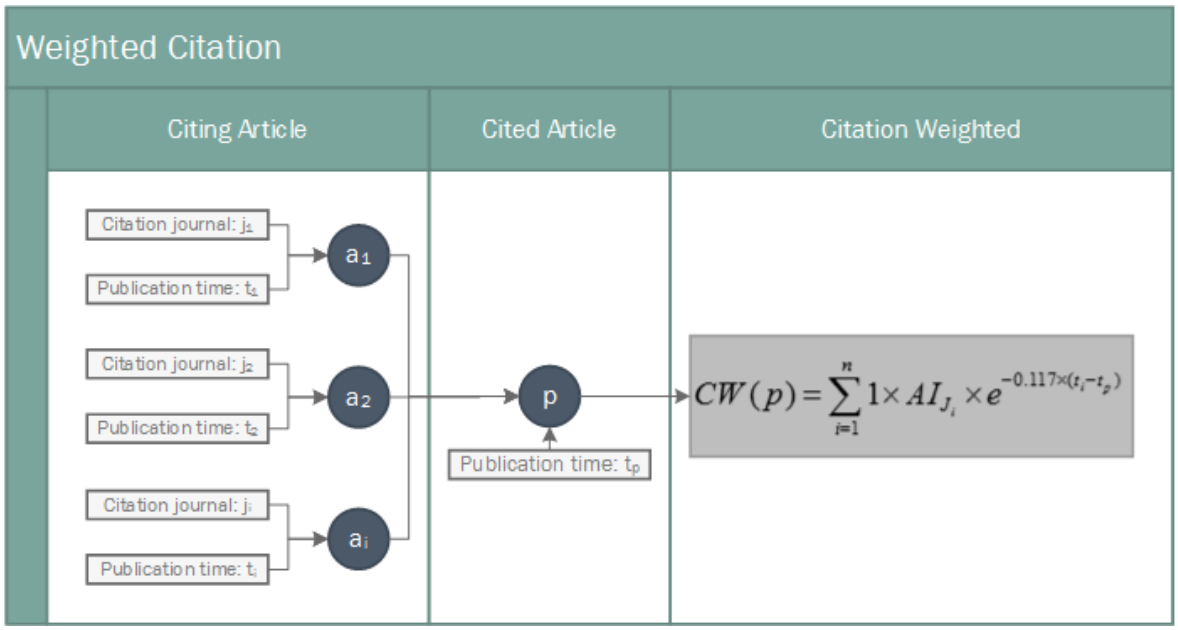
Σχήμα 2.9: Το πλήθος των παραπομπών που λαμβάνει κάθε δημοσίευση σε σχέση με την ηλικία της. Το διάγραμμα προέρχεται από την εργασία [7] και βασίζεται στο πλήθος των παραπομπών των άρθρων που αναρτήθηκαν στην ψηφιακή βιβλιοθήκη JASIST [13] μέχρι και το έτος 2008 (η τιμή 0 στον οριζόντιο άξονα αντιστοιχίζεται στα άρθρα δημοσιευμένα το 2008, η τιμή 1 στα άρθρα δημοσιευμένα το έτος 2007 κ.ο.κ.). Το πλήθος των παραπομπών αντλήθηκε από την εργασία του περιοδικού *Journal Citation Reports* του ιδρύματος Thomson Reuters [10], τεύχος του 2008.

Από τη γραφική παράσταση του Σχήματος 2.9 προκύπτει ότι το μεγαλύτερο πλήθος των παραπομπών προς μία εργασία συγκεντρώνεται 2 χρόνια μετά τη δημοσίευσή της και μειώνεται εκθετικά σε σχέση με τον χρόνο με ρυθμό $f(x) \sim e^{-0.117x}$. Γι' αυτό το λόγο, κατά την εφαρμογή του προτεινόμενου αλγόριθμου, τα βάρη των ακμών πολλαπλασιάζονται με την ποσότητα $e^{-0.117(t_{citation}-t_{publication})}$. Με αυτόν τον τρόπο, οι ακμές που αντιστοιχούν σε παραπομπές που δημιουργήθηκαν πρόσφατα σχετικά με την ημερομηνία δημοσίευσης, θα έχουν μεγαλύτερο βάρος υποδεικνύοντας ότι όταν μια εργασία λαμβάνει γρήγορα παραπομπές, είναι αρκετά σημαντική.

Όσον αφορά τις παλιές εργασίες, τα βάρη των ακμών τους θα ποικίλουν, μιας και έχει περάσει αρκετός καιρός από τη δημοσίευσή τους και η τιμή του $t_{citation}$ πρόκειται να λαμβάνει μεγάλη γκάμα τιμών. Αντίθετα, το σύνολο των ακμών που αντιστοιχούν σε παραπομπές προς τις νεότερες δημοσιεύσεις θα έχει βάρη με μεγάλες τιμές. Με αυτό τον τρόπο, ναι μεν προωθούνται οι νέες δημοσιεύσεις, όπως και στην προηγούμενη ενότητα, αλλά μόνο αυτές

που έχουν λάβει ήδη ένα σύνολο εισερχόμενων ακμών. Αποφεύγουμε έτσι την προώθηση δημοσιεύσεων οι οποίες αν και πρόσφατες δεν είναι σημαντικές.

Στο Σχήμα 2.10 φαίνονται τα βήματα του προτεινόμενου αλγόριθμου:



Σχήμα 2.10: Το μοντέλο του αλγόριθμου κατάταξης σε γράφο με σταθμισμένες ακμές.

Παράδειγμα

Έστω ότι θέλουμε να βρούμε το σκορ κατάταξης της εργασίας D, δημοσιευμένης το 2005, στην οποία παραπέμπουν οι εργασίες A, B και C, δημοσιευμένες το 2005, 2006 και 2007 αντίστοιχα. Για τον υπολογισμό χρησιμοποιείται η ακόλουθη διαδικασία:

- Εντοπίζονται τα περιοδικά στα οποία ανήκουν οι δημοσιεύσεις που παραπέμπουν στη D. Οι δημοσιεύσεις A, B, C ανήκουν στα περιοδικά J_A, J_B, J_C αντίστοιχα.
- Υπολογίζεται το Article Influence score των περιοδικών που είναι δημοσιευμένες οι εργασίες που παραπέμπουν στην D σύμφωνα με τον τύπο 2.10. Το σκορ αυτό για κάθε περιοδικό υπολογίζεται για τη χρονολογία που δημοσιεύτηκε σε αυτό η αντίστοιχη εργασία. Στο παράδειγμά μας το J_A είχε Article Influence score AI_{J_A} το 2005, το J_B είχε AI_{J_B} το 2006 και το J_C είχε AI_{J_C} το 2007.
- Υπολογίζεται το βάρος του κάθε εισερχόμενου συνδέσμου ανάλογα με το χρονικό διάστημα που έχει περάσει από τη δημοσίευση της εργασίας D: $e^{-0.117\times(2005-2005)}$ για την A, $e^{-0.117\times(2006-2005)}$ για τη B και $e^{-0.117\times(2007-2005)}$ για τη C.
- Υπολογίζεται το σκορ κατάταξης της εργασίας D: $e^{-0.117\times(2005-2005)} \times AI_{J_A} + e^{-0.117\times(2006-2005)} \times AI_{J_B} + e^{-0.117\times(2007-2005)} \times AI_{J_C} = AI_{J_A} + 0.89AI_{J_B} + 0.79AI_{J_C}$.

2.4.3 Ο αλγόριθμος FutureRank

Η κατάταξη των επιστημονικών περιοδικών καθίσταται περίπλοκη εξαιτίας του γεγονότος ότι με το πέρασμα του χρόνου ο γράφος των επιστημονικών εργασιών και των παραπομπών τους συνεχώς εμπλουτίζεται.

Ο αλγόριθμος PageRank λειτουργεί σωστά για το σκοπό που έχει σχεδιαστεί, να υπολογίζει δηλαδή το κύρος των εργασιών βασισμένος στο τρέχον δίκτυο των εργασιών και των παραπομπών τους. Παρ' όλα αυτά, δεν λαμβάνει υπόψη ότι πολλές από τις εργασίες πρόκειται να λάβουν αρκετά μεγάλο πλήθος εισερχόμενων συνδέσμων στο μέλλον. Για να είναι λοιπόν έγκυρα τα αποτελέσματα του PageRank, απαιτείται επανυπολογισμός των σκορ στον ανανεωμένο γράφο ανά τακτά χρονικά διαστήματα.

Στην Εργασία [8] ορίστηκε ένας νέος αλγόριθμος κατάταξης των δημοσιεύσεων, που προσπαθεί να προβλέψει το σκορ PageRank που πρόκειται να έχουν οι δημοσιεύσεις στο μέλλον. Με άλλα λόγια, τα σκορ που προκύπτουν από τον αλγόριθμο αυτό προσεγγίζουν τα σκορ PageRank, τα οποία θα υπολογίζονταν πάνω στο γράφο που δημιουργείται μόνο από τις μελλοντικές αναφορές. Πέρα από τον γράφο των δημοσιεύσεων και τις παραπομπές τους, για να προβλεφθούν οι μελλοντικές αναφορές, χρησιμοποιείται το σύνολο των συγγραφέων, οι οποίοι συνδέονται με τις δημοσιεύσεις που έχουν γράψει, καθώς και οι χρονολογίες δημοσίευσης.

Το προτεινόμενο μοντέλο κατάταξης βασίζεται στις παρακάτω παραδοχές:

1. *Σημαντικές επιστημονικές εργασίες δέχονται παραπομπές από σημαντικές επιστημονικές εργασίες.*

2. *Σημαντικές επιστημονικές εργασίες γράφονται από ευυπόληπτους συγγραφείς και οι συγγραφείς είναι ευυπόληπτοι επειδή γράφουν σημαντικές εργασίες. Αυτό δείχνει αλληλοενίσχυση μεταξύ των εργασιών και των συγγραφέων τους.*

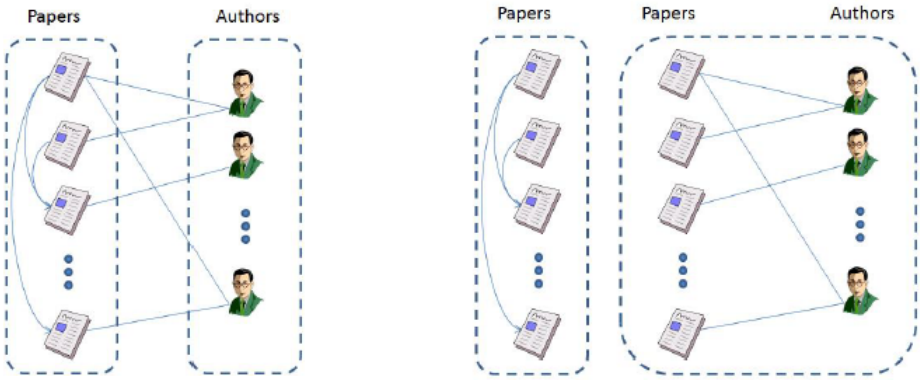
3. *Πρόσφατα δημοσιευμένες εργασίες είναι πιο χρήσιμες ή, με άλλα λόγια, πρόκειται να λάβουν περισσότερες παραπομπές στο άμεσο μέλλον.*

4. *Μεταξύ των παλιότερα δημοσιευμένων εργασιών, πιο σημαντικές είναι αυτές που έχουν λάβει πιο πρόσφατα εισερχόμενους συνδέσμους.*

Η δομή του δικτύου

Το δίκτυο που κατασκευάζεται έχει δύο είδη κόμβων, τις εργασίες και τους συγγραφείς. Έχει επίσης και δύο είδη ακμών, μη-κατευθυνόμενες ακμές μεταξύ των εργασιών και των συγγραφέων τους και κατευθυνόμενες ακμές μεταξύ των εργασιών ανάλογα με τις παραπομπές που δημιουργούνται από τη βιβλιογραφία τους. Για να κατατάξουμε τους κόμβους του δικτύου, το θεωρούμε σαν συνδυασμό δύο δικτύων. Το πρώτο περιέχει μόνο τους κόμβους των εργασιών και τις μεταξύ τους παραπομπές και το δεύτερο περιέχει και τα δύο είδη κόμβων αλλά μόνο τις ακμές που αφορούν τις συνδέσεις εργασιών-συγγραφέων.

Το δίκτυο μοντελοποείται ως εξής:



Σχήμα 2.11: Η δομή του δικτύου που χρησιμοποιείται από τον αλγόριθμο FutureRank.

Αν P σύνολο δημοσιεύσεων και A σύνολο συγγραφέων, ο πίνακας γειτνίασης μεταξύ των δημοσιεύσεων και των παραπομπών τους, $|P| \times |P|$ διαστάσεων είναι ο:

$$M_{i,j}^{C} = \begin{cases} 1 \text{ if } p_i \text{ cites } p_j \\ 0 \text{ otherwise} \end{cases}$$

Για κάθε δημοσίευση p_i που δεν περιέχει παραπομπές προς άλλες εργασίες, $M_{i,j}^{C} = 1$ για όλα τα j. Ο πίνακας γειτνίασης μεταξύ των εργασιών και των συγγραφέων τους, $|P| \times |A|$ διαστάσεων είναι ο:

$$M_{i,j}^{A} = \begin{cases} 1 \text{ if } a_i \text{ is the author of } p_j \\ 0 \text{ otherwise} \end{cases}$$

Ορισμός του αλγόριθμου FutureRank

Από τη στιγμή που οι δύο γράφοι έχουν κοινούς κόμβους, δεν μπορούμε να υπολογίσουμε ξεχωριστά τα σκορ των κατατάξεων για κάθε έναν από αυτούς. Εναλλακτικά, διατρέχουμε ταυτόχρονα και τους δύο γράφους. *Έστω R^P το διάνυσμα με τα τελικά σκορ κατάταξης των δημοσιεύσεων και R^A το διάνυσμα με τα τελικά σκορ κατάταξης των συγγραφέων:*

$$\begin{aligned} R^A &= M^A \cdot R^P \\ R^P &= aM^{C^T} \cdot R^C + \beta(M^{A^T} \cdot R^A) + \gamma R^{time} + (1-a-b-\gamma)[1/n] \end{aligned} \tag{2.15}$$

Οι όροι του παραπάνω τύπου είναι οι εξής:

- $M^{C^T} \cdot R^C$ υπολογίζει το PageRank των δημοσιεύσεων με βάση τον τύπο 2.2.
- $M^{A^T} \cdot R^A$ υπολογίζει τα σκορ κατάταξης των συγγραφέων στο αντίστοιχο δίκτυο.
- R^{Time} αποτελεί ένα «εξατομικευμένο» διάνυσμα του αλγόριθμου PageRank, το οποίο προϋπολογίζεται και αντιπροσωπεύει τις προτιμήσεις του χρήστη. Στη συγκεκριμένη περίπτωση:

  $$R_i^{Time} = e^{-\rho(T_{current} - T_i)}$$

  όπου T_i το έτος δημοσίευσης του p_i και $T_{current}$ το τρέχον έτος.

Το διάνυσμα αυτό λαμβάνει την τιμή που αναφέραμε, εξαιτίας του τύπου που περιγράφει το πλήθος των παραπομπών που λαμβάνει μία δημοσίευση ανά έτος και φαίνεται στο Σχήμα 2.9 της προηγούμενης ενότητας.

Ο αλγόριθμος που περιγράφεται είναι επαναληπτικός. Σε κάθε επανάληψη, οι δημοσιεύσεις μεταφέρουν τα σκορ που έχουν μέχρι εκείνη τη στιγμή στους συγγραφείς τους και αντίστροφα.

Η αρχική τιμή του R_i^P ορίζεται ως $\frac{1}{|P|}$ και η αρχική τιμή του R_i^A είναι $\frac{1}{|A|}$. Η συγκεκριμένη αρχικοποίηση κρατάει κανονικοποιημένα τα σκορ που προκύπτουν από την επαναληπτική διαδικασία αφού το άθροισμα των βαρών $\alpha + \beta + \gamma + (1 - \alpha - \beta - \gamma)$ ισούται με 1.

Υπάρχει μία κατάσταση που δεν περιγράφεται με ξεχωριστό όρο στη συνάρτηση. Αυτή είναι η περίπτωση των σημαντικών εργασιών οι οποίες δεν είναι πρόσφατα δημοσιευμένες αλλά εξακολουθούν να λαμβάνουν παραπομπές από άλλες εργασίες. Ο προτεινόμενος αλγόριθμος πρόκειται να κατατάξει ψηλά αυτές τις εργασίες, γιατί δέχονται παραπομπές και από νέες εργασίες που προωθούνται από τον αλγόριθμο και γι' αυτό το λόγο έχουν μεγάλο σκορ. Το σκορ αυτό το μεταφέρουν και στις εν λόγω εργασίες.

2.4.4 Η εφαρμογή Diana Mirpub

Όπως αναφέρθηκε παραπάνω, η εφαρμογή Diana Mirpub αποτελεί μια μηχανή αναζήτησης επιστημονικών δημοσιεύσεων σχετικών με microRNAs. Στη βάση της είναι καταχωρημένες πάνω από 20.690 δημοσιεύσεις οι οποίες σχετίζονται με 31.984 διαφορετικές λέξεις-κλειδιά που περιγράφουν microRNAs.

Οι συσχετίσεις μεταξύ δημοσιεύσεων και microRNAs παράγονται με τους παρακάτω τρόπους:

Text mining τεχνικές.

Οι συσχετίσεις προέκυψαν εφαρμόζοντας συγκεκριμένους κανόνες στους τίτλους και τις περιλήψεις των άρθρων της MEDLINE και PMC που συγκεντρώθηκαν μέσω της PubMed. Ιδιαίτερη σημασία δόθηκε στην εξέλιξη των καταγραφόμενων δεδομένων για τα microRNAs και ιδιαίτερα στις παραλλαγές των ονομάτων τους.

Η βάση δεδομένων MEDLINE περιέχει δημοσιεύσεις περιοδικών ιατρικού περιεχομένου και αποτελεί μία από τις βάσεις δεδομένων της Εθνικής Βιβλιοθήκης Ιατρικής των Η.Π.Α. (NLM: National Library of Medicine). Είναι άμεσα προσπελάσιμη από την NLM καθώς και από πολυάριθμες μηχανές αναζήτησης. Αποτελεί υποσύνολο της βάσης PubMed και έχει το πλεονέκτημα σε σχέση με τα υπόλοιπα στοιχεία της βάσης αυτής ότι χρησιμοποιεί τις επικεφαλίδες ιατρικού περιεχομένου (MeSH: Medical Subject Headings [16]) της NLM για την ευρετηρίαση των δημοσιεύσεων.

Η βάση δεδομένων PubMed είναι διαθέσιμη από το 1966. Διαθέτει πάνω από 25 εκατομμύρια δημοσιεύσεις, συμπεριλαμβανομένων και αυτών της MEDLINE. Για κάθε δημοσίευση είναι καταχωρημένες και οι παραπομπές που προκύπτουν με τους παρακάτω τρόπους:

- «Παραπομπές σε εξέλιξη»: παραπομπές σε άρθρα που δεν έχουν ελεγχθεί ακόμα.
- Παραπομπές σε άρθρα εκτός πεδίου (κυρίως γενικών επιστημών) που δεν δεικτοδοτούνται με βάση τις επικεφαλίδες ιατρικού περιεχομένου (MeSH) της NLM.
- «Παραπομπές πριν από την εκτύπωση»: Παραπομπές που υπήρχαν σε ένα άρθρο πριν την τελική δημοσίευσή του.
- Παραπομπές προς τα άρθρα ενός περιοδικού που προηγούνται της ημερομηνίας που αυτό καταχωρήθηκε στη MEDLINE.
- Παραπομπές σε άρθρα δημοσιευμένα πριν το 1966, οι οποίες δεν συμπεριλαμβάνονται στη MEDLINE διότι δεν έχουν δεικτοδοτηθεί με βάση τη MeSH τεχνική.
- Παραπομπές σε κάποια επιπλέον περιοδικά επιστημών ζωής, τα οποία είναι καταχωρημένα στη βάση PMC (PubMed Central) .
- Παραπομπές σε χειρόγραφα συγκεκριμένων συγγραφέων.
- Παραπομπές προς την πλειοψηφία των βιβλίων που είναι διαθέσιμα στη NCBI βιβλιοθήκη [19].

Η βάση δεδομένων PMC (PubMed Central) αποτελεί αρχείο με το πλήρες κείμενο περιοδικών με άρθρα στο πεδίο της βιοϊατρικής και των βιοεπιστημών. Τα άρθρα του αρχείου PMC αποτελούν υποσύνολο των άρθρων της PubMed.

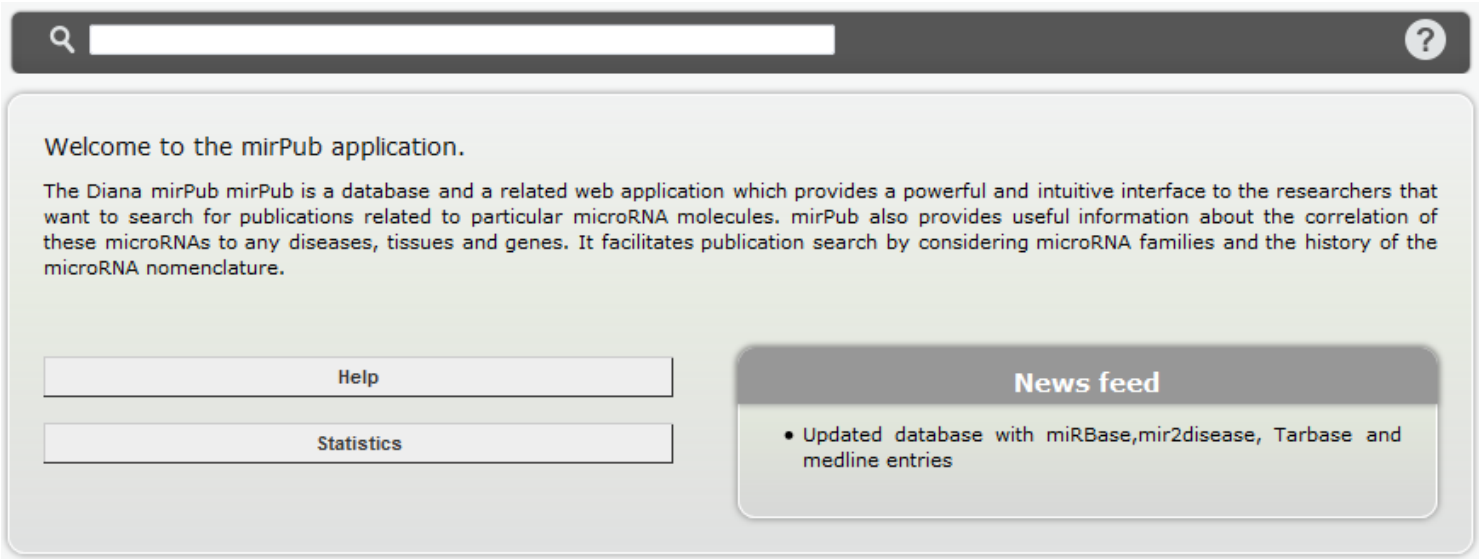
miRBase, Tarbase v.6.0, mir2disease.

Ένα σύνολο των συσχετίσεων συλλέχθηκε από τις βάσεις δεδομένων miRBase, Tarbase v.6.0 και mir2disease

Συσχετίσεις παραγόμενες από τους χρήστες.

Η εφαρμογή παρέχει τη δυνατότητα στους χρήστες να συμβάλλουν στα περιεχόμενά της, προτείνοντας συσχετίσεις microRNAs - δημοσιεύσεων που δεν υπάρχουν ακόμα.

Η αναζήτηση δημοσιεύσεων με βάση ένα microRNA πραγματοποιείται εισάγοντας το όνομα του microRNA στη φόρμα αναζήτησης της διεπαφής της εφαρμογής. Οι δημοσιεύσεις που επιστρέφονται σαν αποτέλεσμα της αναζήτησης, δεν περιέχουν μόνο τη λέξη-κλειδί που εισήγαγε ο χρήστης αλλά και αυτές που αποτελούν παραλλαγές του ονόματος του αντίστοιχου microRNA καθώς και προγενέστερα ή μεταγενέστερα ονόματά του [32]. Στη συνέχεια παρέχεται στον χρήστη η δυνατότητα να φιλτράρει τα αποτελέσματα επιλέγοντας μία λέξη-κλειδί από το παραπάνω σύνολο.



Σχήμα 2.12: Τμήμα της αρχικής σελίδας της εφαρμογής mirPub

Κεφάλαιο 3

Σχεδίαση και υλοποίηση αλγορίθμων κατάταξης

Στο παρόν κεφάλαιο, περιγράφονται όλοι οι αλγόριθμοι που υλοποιήθηκαν στα πλαίσια της διπλωματικής, οι οποίοι βασίζονται στους αλγόριθμους των επιστημονικών εργασιών που παρουσιάζονται στο Κεφάλαιο 2. Στη συνέχεια, αναλύεται το σύνολο των χαρακτηριστικών που πρέπει να πληροί κάθε αλγόριθμος, καθώς και τα προγραμματιστικά εργαλεία που χρησιμοποιήθηκαν. Τέλος παρουσιάζονται τα τελικά προγράμματα.

3.1 Αλγόριθμοι που υλοποιήθηκαν

3.1.1 PageRank

Υλοποιήθηκε ο κλασικός αλγόριθμος PageRank που περιγράφεται στο Κεφάλαιο 2. Η παράμετρος α, η οποία σταθμίζει το βάρος των μεταβάσεων από συνδέσμους, έναντι των τυχαίων μεταβάσεων τέθηκε ίση με 0.5. Η τιμή της παραμέτρου αυτής που χρησιμοποιείται στην περίπτωση που θέλουμε να κατατάξουμε ιστοσελίδες του Διαδικτύου είναι 0.85. Σε αυτή την περίπτωση, υιοθετούμε την εμπειρική παρατήρηση ότι ένας ερευνητής ακολουθεί λιγότερους συνδέσμους (αναφορές) απ' ότι ένας χρήστης του Διαδικτύου και χρησιμοποιήσαμε την τιμή που προτείνεται από τη βιβλιογραφία [22]. Ο χρήστης του Διαδικτύου λοιπόν, με βάση τις παραπάνω τιμές, ακολουθεί 6 συνδέσμους κατά μέσο όρο, ενώ ο ερευνητής 2.

3.1.2 AdvRecPubs: κατάταξη με βάση την προώθηση των νέων δημοσιεύσεων

Πραγματοποιήθηκε μια πρώτη υλοποίηση του αλγόριθμου που προτάθηκε από την Εργασία [29] και περιγράφεται στην Ενότητα 2.4.1. Στην υλοποίηση αυτήν, ικανοποιούνται ο πρώτος και τρίτος στόχος όπως αναφέρονται στην προηγούμενη ενότητα. Συνεπώς, στον τύπο του Pagerank (2.4), το διάνυσμα που όλα τα στοιχεία του ισούνται με τη μονάδα, αντικαταστάθηκε

με το **w** που περιγράφεται παρακάτω:

$$\mathbf{w}[i] = \frac{\frac{e^{(\mathbf{y}[i]-c)/b}}{b}}{\sum_{j=1}^{N} \frac{e^{(\mathbf{y}[j]-c)/b}}{b}} \tag{3.1}$$

όπου **y**[i] το έτος δημοσίευσης του άρθρου i και c το τρέχον έτος και b σταθερά.

Έτσι ώστε ο τύπος του να διαμορφωθεί ως εξής:

$$\mathbf{p_{i+1}} = a(\mathbf{G} + \mathbf{w} \cdot \mathbf{d}^T)\mathbf{p_i} + (1-a)\mathbf{w} \tag{3.2}$$

Αυτό διαισθητικά σημαίνει ότι ο τυχαίος ερευνητής όταν σταματήσει να ακολουθεί τις αναφορές των δημοσιεύσεων, υπάρχει μεγαλύτερη πιθανότητα να επανεκκινήσει την αναζήτησή του από μία σχετικά πρόσφατη δημοσίευση. Επίσης οι "τεχνητές ακμές" που δημιουργήθηκαν για τις δημοσιεύσεις χωρίς αναφορές δείχνουν κυρίως νέες δημοσιεύσεις. Σε αυτήν όπως και στην προηγούμενη περίπτωση $\alpha = 0.5$. Θα αναφερόμαστε σε αυτόν τον αλγόριθμο με το όνομα **AdvRecPubs** (Advocating Recent Publications).

3.1.3 Κατάταξη με βάση τη δημοτικότητα/κύρος του αντίστοιχου περιοδικού δημοσίευσης

Πραγματοποιήθηκε μια δεύτερη προσέγγιση του αλγόριθμου που περιγράφεται στην εργασία [29] με τη διαφορά ότι σε αυτή την περίπτωση ικανοποιούνται ο πρώτος και δεύτερος στόχος όπως τίθενται στην εργασία. Συνεπώς, ο τύπος του υλοποιημένου αλγόριθμου είναι ο ίδιος ((3.2)) με τη διαφορά ότι το διάνυσμα **w** διαμορφώνεται από τη δημοτικότητα ή το κύρος του περιοδικού στο οποίο είναι δημοσιευμένο κάθε άρθρο.

Αυτό διαισθητικά σημαίνει ότι ο τυχαίος ερευνητής όταν σταματήσει να ακολουθεί τις αναφορές των δημοσιεύσεων, υπάρχει μεγαλύτερη πιθανότητα να επανεκκινήσει την αναζήτησή του από ένα άρθρο που είναι δημοσιευμένο σε δημοφιλές-έγκυρο περιοδικό. Επίσης οι τεχνητές ακμές που δημιουργήθηκαν για τις δημοσιεύσεις χωρίς αναφορές δείχνουν κυρίως άρθρα δημοσιευμένα σε δημοφιλή-έγκυρα περιοδικά. Σε αυτήν όπως και στην προηγούμενη περίπτωση $\alpha = 0.5$.

Υπολογίσαμε τη δημοτικότητα των περιοδικών με 3 διαφορετικούς τρόπους, όπως παρουσιάζονται στις επόμενες ενότητες και το κύρος των περιοδικών με τη χρήση του Article Influence Score. Θα αναφερόμαστε στους αλγόριθμους που χρησιμοποιούν το Impact Factor με το όνομα **RankWithIF**, και σε αυτόν που χρησιμοποιεί το Article Influence Score με το όνομα **RankWithAI**.

3.1.3.1 RankWithIF1: Impact factor υπολογισμένο σε διάστημα 2 ετών

Η τιμή του impact factor κάθε περιοδικού ισούται με το πλήθος των αναφορών των άρθρων που δημοσιεύτηκαν το τρέχον έτος που δείχνουν προς τα άρθρα που δημοσιεύτηκαν

τα 2 προηγούμενα έτη στο περιοδικό αυτό, προς στο πλήθος των άρθρων αυτών. Επομένως το διάνυσμα **w** διαμορφώνεται ώς εξής:

$$\mathbf{w}[i] = \frac{\mathbf{IF}(v_i, y : 2)}{\sum_{j=1}^{N} \mathbf{IF}(v_j, y : 2)} \tag{3.3}$$

όπου v_i το περιοδικό στο οποίο είναι δημοσιευμένο το άρθρο i, y το τρέχον έτος και $\mathbf{IF}(v_i, y : t)$ ο τύπος (2.6).

3.1.3.2 RankWithIF2: Impact factor υπολογισμένο σε διάστημα 5 ετών

Η τιμή του impact factor κάθε περιοδικού ισούται με το πλήθος των αναφορών των άρθρων που δημοσιεύτηκαν το τρέχον έτος που δείχνουν προς τα άρθρα που δημοσιεύτηκαν τα 5 προηγούμενα έτη στο περιοδικό αυτό, προς στο πλήθος των άρθρων αυτών. Επομένως το διάνυσμα **w** διαμορφώνεται ως εξής:

$$\mathbf{w}[i] = \frac{\mathbf{IF}(v_i, y : 5)}{\sum_{j=1}^{N} \mathbf{IF}(v_j, y : 5)} \tag{3.4}$$

όπου v_i το περιοδικό στο οποίο είναι δημοσιευμένο το άρθρο i, y το τρέχον έτος και $\mathbf{IF}(v_i, y : t)$ ο τύπος (2.6).

3.1.3.3 RankWithIF3: Impact factor με μεταβαλλόμενη την τιμή τρέχοντος έτους

Το impact factor του περιοδικού ενός άρθρου είναι υπολογισμένο το έτος δημοσίευσης του άρθρου αυτού. Με αυτό τον τρόπο η τιμή του διανύσματος **w** διαφέρει ακόμα και για άρθρα που είναι δημοσιευμένα στο ίδιο περιοδικό, αφού ο υπολογισμός του impact factor δεν γίνεται στο τρέχον αλλά σε μεταβαλλόμενο έτος. Επομένως το διάνυσμα **w** διαμορφώνεται ως εξής:

$$\mathbf{w}[i] = \frac{\mathbf{IF}(v_i, y_i : 5)}{\sum_{j=1}^{N} \mathbf{IF}(v_j, y_j : 5)} \tag{3.5}$$

όπου v_i το περιοδικό στο οποίο είναι δημοσιευμένο το άρθρο i, y_i το έτος δημοσίευσης του άρθρου i και $\mathbf{IF}(v_i, y : t)$ ο τύπος (2.6).

3.1.3.4 RankWithAI: Article Influence score

Το διάνυσμα **w** διαμορφώνεται ώς εξής:

$$\mathbf{w}[i] = \frac{\mathbf{AI}[v_i]}{\sum_{j=1}^{N} \mathbf{AI}[v_j]} \tag{3.6}$$

όπου v_i το περιοδικό στο οποίο είναι δημοσιευμένο το άρθρο i και $\mathbf{AI}[v_i]$ ο τύπος 2.10.

Το όνομα του συγκεκριμένου αλγόριθμου καθορίζεται από το χρονικό διάστημα στο οποίο υπολογίζεται το EigenFactor (RankWithAI1: υπολογισμένο σε διάστημα 2 ετών RankWithAI2: υπολογισμένο σε διάστημα 5 ετών.)

3.1.4 AdvRecPubs-RankWithIF: αλγόριθμος κατάταξης με βάση την προώθηση των νέων δημοσιεύσεων και το impact factor του αντίστοιχου περιοδικού

Υλοποιήθηκε ο αλγόριθμος που περιγράφεται στην Ενότητα 2.4.1. Το impact factor αντικαταστάθηκε με όλες τις υλοποιημένες μετρικές δημοτικότητας των περιοδικών και επιλέχτηκε αυτή για την οποία προκύπτουν τα καλύτερα αποτελέσματα. Η σύγκριση και εκτίμηση των αποτελεσμάτων πραγματοποιείται στο επόμενο κεφάλαιο. Θα αναφερόμαστε στον συγκεκριμένο αλγόριθμο με το όνομα **AdvRecPubs-RankWithIF**. Ανάλογα με τη σταθερά b που χρησιμοποιείται για την προώθηση των δημοσιεύσεων και τον τρόπο υπολογισμού του impact factor διαμορφώνεται και το όνομά του (π.χ. όταν $b = 3$ και το impcat factor είναι υπολογισμένο σε διάστημα 2 ετών τότε το όνομα του αλγόριθμου είναι AdvRecPubs3-RankWithIF1).

3.1.5 Weighted Citations: αλγόριθμος κατάταξης σε γράφο με σταθμισμένες ακμές

Υλοποιήθηκε ο αλγόριθμος που περιγράφεται στην Ενότητα 2.4.2. Θα αναφερόμαστε στον συγκεκριμένο αλγόριθμο με το όνομα **Weighted Citations**. Ανάλογα με τον τρόπο υπολογισμού του Eigenfactor διαμορφώνεται και το όνομα του αλγόριθμου (Weighted Citations1: το EigenFactor είναι υπολογισμένο σε διάστημα 2 ετών, Weighted Citations2: το EigenFactor είναι υπολογισμένο σε διάστημα 5 ετών).

3.1.6 FutureRankLike: παραλλαγή του αλγόριθμου FutureRank

Δημιουργήσαμε μία παραλλαγή του αλγόριθμου FutureRank. Συγκεκριμένα χρησιμοποιήσαμε την ιδέα της προσθήκης ενός επιπλέον όρου στον αλγόριθμο του Pagerank. Με αυτό τον τρόπο ο τύπος του Pagerank (2.4) διαμορφώνεται ως εξής:

$$\mathbf{p_{i+1}} = a(\mathbf{G} + \mathbf{z} \cdot \mathbf{d}^T)\mathbf{p_i} + b\mathbf{v} + (1 - a - b)\mathbf{w} \tag{3.7}$$

όπου $\mathbf{G}$ ο πίνακας γειτνίασης του γράφου, $\mathbf{v}$ κανονικοποιημένο διάνυσμα με τιμή ίση με τη μετρική δημοτικότητας/κύρους του αντίστοιχου περιοδικού, $\mathbf{w}$ το διάνυσμα που περιγράφεται από τον τύπο 3.1, $\mathbf{z}$ όπου $\mathbf{z}_i = \mathbf{v}_i * \mathbf{w}_i$, $\mathbf{d}$ διάνυσμα με τιμή ίση με 1 αν ο αντίστοιχος κόμβος δεν έχει εξερχόμενες ακμές και 0 αν έχει και a,b βάρη για τα οποία ισχύει $0 \leq a, b \leq 1$.

Με αυτόν τον τρόπο ο τυχαίος ερευνητής ακολουθεί με πιθανότητα a τον γράφο των αναφορών των δημοσιεύσεων, με πιθανότητα b ξεκινά την αναζήτηση από άρθρα δημοσιευμένα σε δημοφιλή περιοδικά και με πιθανότητα $1 - a - b$ ξεκινά την αναζήτηση από πρόσφατα δημοσιευμένα άρθρα. Επίσης οι "τεχνητές ακμές" που δημιουργήθηκαν για τις δημοσιεύσεις χωρίς αναφορές δείχνουν κυρίως νέες δημοσιεύσεις, δημοσιευμένες σε δημοφιλή περιοδικά.

Σαν μετρική δημοτικότητας-κύρους των περιοδικών χρησιμοποιήσαμε όλους τους τύπους που υλοποιήσαμε στις προηγούμενες ενότητες δηλαδή το Impact factor των περιοδικών υπολογισμένο σε διάστημα 2 (3.1.3.1), 5 ετών (3.1.3.2) και με μεταβαλλόμενη τιμή τρέχοντος έτους (3.1.3.3) και το Article Influence score των περιοδικών (3.1.3.4). Τέλος καταλήξαμε σε αυτόν για τον οποίο προκύπτουν τα καλύτερα αποτελέσματα. Θα αναφερόμαστε στον συγκεκριμένο αλγόριθμο με το όνομα **FutureRankLike**. Ανάλογα με τη σταθερά b που χρησιμοποιείται για την προώθηση των δημοσιεύσεων και τον τρόπο υπολογισμού του impact factor διαμορφώνεται και το όνομά του (π.χ. όταν $b = 3$ και το impcat factor είναι υπολογισμένο σε διάστημα 2 ετών τότε το όνομα του αλγόριθμου είναι FutureRankLike3.1).

3.1.7 Συνδυαστικοι αλγόριθμοι των παραπάνω υλοποιήσεων

3.1.7.1 Weighted Citations in PageRank: αποτελέσματα αλγόριθμου σαν είσοδος του Pagerank

Ο αλγόριθμος που περιγράφεται στην Ενότητα 2.4.2 εφαρμόζει παραπλήσιο αλγόριθμο με αυτόν του PageRank, τον Eigenfactor, στο δίκτυο των περιοδικών και όχι στο δίκτυο των δημοσιεύσεων. Η τελική κατάταξη επομένως, γίνεται κυρίως με βάση το κύρος των αντίστοιχων περιοδικών και το ποσοτικό και όχι ποιοτικό πλήθος των αναφορών προς μία δημοσίευση. Γι' αυτό το λόγο, θέσαμε το διάνυσμα παραμετροποίησης του αλγόριθμου Pagerank $\mathbf{w}$, ίσο με το διάνυσμα των αποτελεσμάτων του παραπάνω αλγόριθμου. Θα αναφερόμαστε στον συγκεκριμένο αλγόριθμο με το όνομα **Weighted Citations in PageRank**.

3.1.7.2 Ποσοστιαία χρήση των αποτελεσμάτων των αλγορίθμων

Η αδυναμία του αλγόριθμου Pagerank να κατατάξει σωστά τις δημοσιεύσεις εστιάζεται στις νέες δημοσιεύσεις και όχι στο σύνολό τους. Τα πρόσφατα άρθρα είναι αυτά που δεν έχουν "προλάβει" να αποκτήσουν επαρκή αριθμό εισερχόμενων ακμών ώστε να πραγματοποιηθεί η κατάταξη με βάση τη δημοτικότητά τους. Με άλλα λόγια, για ένα νέο άρθρο είναι χρήσιμο να "μαντέψουμε" τη σπουδαιότητά του με βάση άλλους παράγοντες όπως το κύρος του περιοδικού στο οποίο είναι δημοσιευμένο. Πραγματοποιώντας όμως το ίδιο και για τα παλιά άρθρα μπορεί να οδηγηθούμε σε λανθασμένα αποτελέσματα. Ένα παλιό άρθρο έχει αποδείξει αν είναι αρκετά σημαντικό ακόμα και όταν δεν είναι δημοσιευμένο σε γνωστό περιοδικό, όπως και το αντίθετο, αν είναι δηλαδή αδιάφορο, ακόμα και όταν είναι δημοσιευμένο σε γνωστό περιοδικό.

Για τους παραπάνω λόγους, ο τελευταίος αλγόριθμος που υλοποιήθηκε αποτελεί συνδυασμό του Pagerank κι ενός από τους αλγορίθμους 3.1.2,3.1.3,3.1.4,3.1.5 και 3.1.6. Συγκεκριμένα, για τις εργασίες που είναι δημοσιευμένες το τρέχον έτος και ένα έτος πριν, τα σκορ κατάταξης δίνονται εξ' ολοκλήρου από τον υλοποιημένο αλγόριθμο που εξετάζουμε κάθε φορά, για τις εργασίες που είναι δημοσιευμένες 2 και 3 έτη πριν το τρέχον έτος, τα σκορ αποτελούνται από το άθροισμα του 5% του PageRank και του 95% του υλοποιημένου αλγόριθμου, για τις εργασίες που είναι δημοσιευμένες τα 4 και 5 έτη πριν το τρέχον έτος, τα σκορ είναι το 10% του PageRank συν το 90% του υλοποιημένου αλγόριθμου, κ.ο.κ..

3.2 Ανάλυση απαιτήσεων

Στόχος είναι να δημιουργηθεί το πρόγραμμα που υλοποιεί όλους τους αλγόριθμους που αναφέρθηκαν παραπάνω. Η κάθε συνάρτηση δέχεται σαν είσοδο μία λίστα με τα αναγνωριστικά των δημοσιεύσεων στη βάση PubMed (pmids), επιστρέφει την αντίστοιχη κατάταξη των δημοσιεύσεων και πληροί τα παρακάτω χαρακτηριστικά.

Συλλογή δεδομένων αυτόματα από το Διαδίκτυο

Για τη δημιουργία της εισόδου των αλγόριθμων απαιτείται κατ' αρχάς η συλλογή των παραπομπών όλων των δημοσιεύσεων. Σε όλες τις περιπτώσεις πέρα απο το Pagerank είναι απαραίτητο και το σύνολο των χρονολογιών και των περιοδικών δημοσίευσης των άρθρων. Η πληροφορία αυτή μπορεί να αντληθεί από τη βάση Pubmed μέσω του E-utilities API της μηχανής αναζήτησης Entrez που περιγράφεται στην Ενότητα 3.3.1.

Επεξεργασία δεδομένων

Τα δεδομένα που συγκεντρώθηκαν πρέπει να επεξεργαστούν κατάλληλα ώστε να δημιουργηθεί η είσοδος κάθε αλγόριθμου. Συγκεκριμένα, με τη χρήση των αναφορών των δημοσιεύσεων πρέπει να κατασκευαστεί ένας γράφος με κόμβους τις δημοσιεύσεις της εισόδου και ακμές τις μεταξύ τους αναφορές. Είναι προτιμότερο ο γράφος αυτός να απεικονίζεται σαν μία δομή δεδομένων που σε κάθε κόμβο αντιστοιχεί μια λίστα με τις ακμές του, παρά σαν πίνακας γειτνίασης μιας και ο γράφος μας είναι αραιός κι έτσι θα δεσμεύαμε πολύ περισσότερη μνήμη.

Υλοποίηση αλγόριθμων

Οι περισσότεροι αλγόριθμοι αποτελούν μια παραμετροποιημένη εκδοχή του αλγόριθμου Pagerank. Επομένως κρίνεται απαραίτητη η υλοποίηση του αλγόριθμου αυτού, με δυνατότητα παραμετροποίησης, ώστε να εισάγεται κάθε φορά μαζί με το γράφο και το κατάλληλο διάνυσμα **w** και να παράγονται τα επιθυμητά αποτελέσματα. Για τον υπολογισμό του διανύσματος **w** πρέπει να υλοποιηθούν ξεχωριστές συναρτήσεις έτσι ώστε να είναι επαναχρησιμοποιήσιμες από τους διάφορους αλγόριθμους.

Για την υλοποίηση του αλγόριθμου **FutureRankLike**, αρκεί να τροποποιήσουμε την παραπάνω συνάρτηση προσθέτοντας έναν επιπλέον όρο στον υπολογισμό του αλγόριθμου.

Για τον υπολογισμό των αποτελεσμάτων του αλγόριθμου **Weighted Citations**, πρέπει η δομή δεδομένων που χρησιμοποιήσαμε για την αποθήκευση του γράφου να μας παρέχει τη δυνατότητα να εισάγουμε βάρη στις ακμές του.

Αποθήκευση δεδομένων εισόδου και αποτελεσμάτων

Ο όγκος των δεδομένων που συλλέγονται είναι αρκετά μεγάλος όπως και ο χρόνος άντλησής τους. Δεδομένου ότι το σύνολο των δημοσιεύσεων της βάσης δεν αλλάζει συχνά, και αν αλλάξει προστίθενται νέες δημοσιεύσεις και δεν διαγράφονται οι παλιές, συμπεραίνουμε ότι θα ήταν χρήσιμη η αποθήκευση του γράφου που δημιουργήσαμε καθώς και των διανυσμάτων με τα έτη και τα περιοδικά δημοσίευσης για την εύκολη επα-

ναχρησιμοποίηση τους. Χρήσιμη θα ήταν και η αποθήκευση των αποτελεσμάτων των αλγορίθμων για την εύκολη επεξεργασία τους και την εξαγωγή συμπερασμάτων.

3.3 Προγραμματιστικά εργαλεία

3.3.1 E-utilities

Η PubMed αποτελεί μια βάση δεδομένων που περιέχει παραπομπές και περιλήψεις βιβλιογραφίας σχετικής με τις επιστήμες της βιολογίας και της βιοϊατρικής. Η βιβλιογραφία αυτή προέρχεται από τη βάση δεδομένων MEDLINE. Η Pubmed παρέχει επίσης συνδέσμους με τα πλήρη κείμενα των δημοσιεύσεων όταν αυτά είναι διαθέσιμα στη βάση δεδομένων Pubmed Central (PMC) ή σε κάποια ιστοσελίδα.

Η βάση δεδομένων Pubmed όπως και η Pubmed Central ενσωματώνονται στη μηχανή αναζήτησης Entrez η οποία παρέχει πρόσβαση σε μεγάλο πλήθος βάσεων δεδομένων που περιέχουν πληροφορίες σχετικές με τη βιοϊατρική και τη γενετική. Η μηχανή αναζήτησης Entrez αποτελεί μέρος της ιστοσελίδας του Εθνικού Κέντρου Πληροφοριών Βιοτεχνολογίας των Η.Π.Α. ([23]).

Τα E-utilities είναι server-side προγράμματα της μηχανής αναζήτησης Entrez που παρέχουν μία διεπαφή για την πρόσβαση στα δεδομένα των βάσεών της. Κάθε πρόγραμμα πραγματοποιεί διαφορετική λειτουργία σχετική με συλλογή δεδομένων και χρησιμοποιείται πραγματοποιώντας HTTP κλήση στο αντίστοιχο URL το οποίο έχει συγκεκριμένη δομή ώστε οι HTTP παράμετροι να αντιστοιχίζονται στις παραμέτρους της εισόδου κάθε εργαλείου. Παρακάτω παρουσιάζονται οι λειτουργίες και παραδείγματα εισόδου/εξόδου των δύο e-utils που χρησιμοποιήθηκαν στα πλαίσια της διπλωματικής. Για μια πλήρη περιγραφή των δυνατοτήτων του εργαλείου E-utils, ο αναγνώστης παραπέμπεται στο
http://www.ncbi.nlm.nih.gov/books/NBK25499/[6].

3.3.1.1 E-Fetch

URL

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi

Λειτουργίες

Επιστρέφει εγγραφές από τη βάση σχετικές με τα αναγνωριστικά (ids) που έχουν δοθεί στην είσοδο.

Επιστρέφει εγγραφές από τη βάση σχετικές με τα αναγνωριστικά (ids) που έχουμε προηγουμένως ανεβάσει στον Entrez History server.

Απαιτούμενες HTTP παράμετροι

db Η βάση από την οποία ζητάται να συγκεντρωθούν τα δεδομένα.

id Λίστα με τα αναγνωριστικά των δημοσιεύσεων (ids) για τις οποίες ζητάται να συγκεντρωθούν δεδομένα από τη βάση.

Προεραιτικές παράμετροι που χρησιμοποιήθηκαν

retmode Η ζητούμενη μορφή των δεδομένων εξόδου.

Παράδειγμα κλήσης του E-Fetch API

HTTP Request

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?
id=1416878&db=pubmed&retmode=xml

Μέρος HTTP Response

```
1  <?xml version="1.0"?>
2  <!DOCTYPE PubmedArticleSet PUBLIC "-//NLM//DTD PubMedArticle, 1st January
   2015//EN" "http://www.ncbi.nlm.nih.gov/corehtml/query/DTD/pubmed_150101.dtd">
3  <PubmedArticleSet>
4    <PubmedArticle>
5      <MedlineCitation Owner="NLM" Status="MEDLINE">
6        <PMID Version="1">1416878</PMID>
7        <DateCreated>
8          <Year>1992</Year>
9        </DateCreated>
10       <Article PubModel="Print">
11         <Journal>
12           <ISSN IssnType="Print">0066-4804</ISSN>
13           <Title>Antimicrobial agents and chemotherapy</Title>
14           <ISOAbbreviation>Antimicrob. Agents Chemother.</ISOAbbreviation>
15         </Journal>
16         <ArticleTitle>Detection of extended-spectrum beta-lactamases in members of
           the family Enterobacteriaceae: comparison of the double-disk and
           three-dimensional tests.</ArticleTitle>
17       </Article>
18       <CommentsCorrectionsList>
19         <CommentsCorrections RefType="Cites">
20           <RefSource>Antimicrob Agents Chemother. 1989
             Nov;33(11):1915-20</RefSource>
21           <PMID Version="1">2558614</PMID>
22         </CommentsCorrections>
23         <CommentsCorrections RefType="Cites">
24           <RefSource>Infection. 1990 Sep-Oct;18(5):294-8</RefSource>
25           <PMID Version="1">2276823</PMID>
26         </CommentsCorrections>
27           .
28           .
29         <CommentsCorrections RefType="ErratumIn">
30           <RefSource>Antimicrob Agents Chemother 1992 Nov;36(11):2575</RefSource>
31         </CommentsCorrections>
32       </CommentsCorrectionsList>
33       <OtherID Source="NLM">PMC192203</OtherID>
34     </MedlineCitation>
35   </PubmedArticle>
36 </PubmedArticleSet>
```

3.3.1.2 E-Link

URL

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi

Λειτουργίες

Επιστρέφει μία λίστα με τις συνδέσεις [1] των δημοσιεύσεων της εισόδου, που ανήκουν είτε στην ίδια είτε σε διαφορετική βάση δεδομένων.

Επιστρέφει μία λίστα με τις συνδέσεις των δημοσιεύσεων της εισόδου, που ανήκουν στην ίδια βάση δεδομένων και ικανοποιούν κάποιο query της μηχανής αναζήτησης Entrez.

Ελέγχει για την ύπαρξη συνδέσεων για ένα σύνολο δημοσιεύσεων μέσα στην ίδια βάση.

Επιστρέφει μία λίστα με τις συνδέσεις ενός συγκεκριμένου άρθρου.

Επιστρέφει μία λίστα με τα URL των παραπομπών των δημοσιεύσεων της εισόδου, καθώς και πληροφορίες για αυτές.

Επιστρέφει μία λίστα με τις δημοσιεύσεις για τις οποίες αυτές της εισόδου αποτελούν εξερχόμενες συνδέσεις.

Απαιτούμενες HTTP παράμετροι

db Η βάση από την οποία ζητάται να συγκεντρωθούν τα δεδομένα.

dbfrom Η βάση από την οποία προέρχονται τα δεδομένα της εισόδου.

cmd Προσδιορίζει ποια λειτουργία από αυτές που παρέχει το Elink ζητάται να πραγματοποιηθεί. Η προεπιλεγμένη τιμή την οποία και χρησιμοποιήσαμε ισούται με neighbor και χρησιμοποιείται για να επιστραφούν όλες οι συνδέσεις της εισόδου.

id Λίστα με τα αναγνωριστικά των δημοσιεύσεων (ids) για τις οποίες ζητάται να συγκεντρωθούν δεδομένα από τη βάση.

[1] Με τον όρο συνδέσεις, εννοούμε τις παραπομπές που εντοπίζονται στη βιβλιογραφία των εργασιών, τους υπερσυνδέσμους που εντοπίζονται στο κείμενό τους κ.τ.λ.. Τα UIDs των άρθρων που επιστρέφονται, αφορούν είτε σε εισερχόμενες (τα άρθρα που παραπέμπουν προς τα άρθρα του εξεταζόμενου σύνολου) είτε σε εξερχόμενες συνδέσεις (τα άρθρα στα οποία παραπέμπουν τα άρθρα του εξεταζόμενου συνόλου).

Παράδειγμα κλήσης του E-Link API

HTTP Request

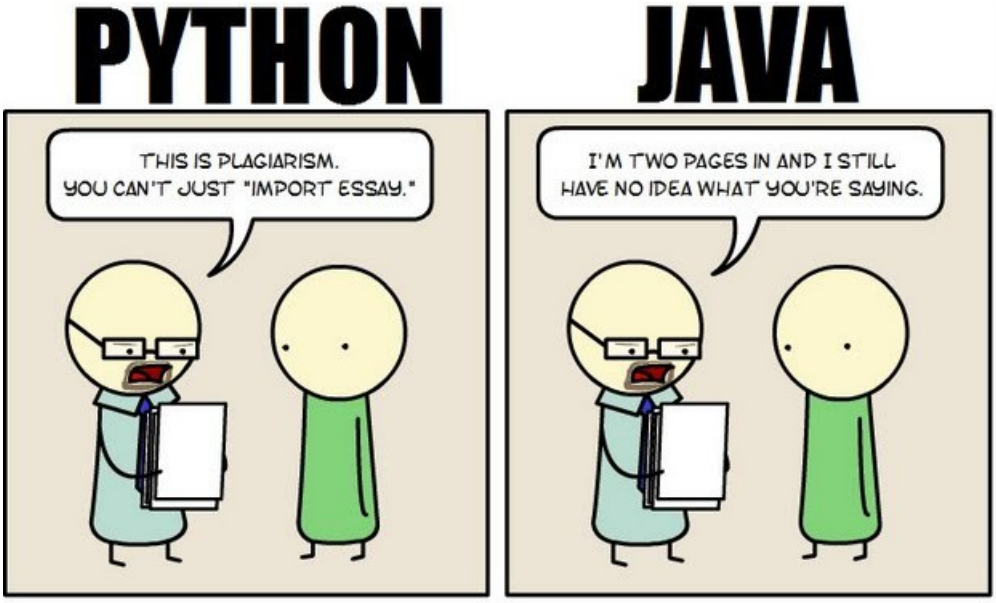
http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?id=1416878
&db=pubmed&dbfrom=pubmed &cmd=neighbor&retmode=xml

Μέρος HTTP Response

```
<?xml version="1.0"?>
<!DOCTYPE eLinkResult PUBLIC "-//NLM//DTD eLinkResult, 23 November 2010//EN"
"http://www.ncbi.nlm.nih.gov/entrez/query/DTD/eLink_101123.dtd">
<eLinkResult>
  <LinkSet>
    <DbFrom>pubmed</DbFrom>
    <IdList>
      <Id>1416878</Id>
    </IdList>
    <LinkSetDb>
      <DbTo>pubmed</DbTo>
      <LinkName>pubmed_pubmed</LinkName>
    </LinkSetDb>
      .
      .
    <LinkSetDb>
      <DbTo>pubmed</DbTo>
      <LinkName>pubmed_pubmed_refs</LinkName>
        <Link><Id>2276823</Id></Link>
        <Link><Id>2197339</Id></Link>
        <Link><Id>2193623</Id></Link>
    </LinkSetDb>
    <LinkSetDb>
      <DbTo>pubmed</DbTo>
      <LinkName>pubmed_pubmed_reviews</LinkName>
    </LinkSetDb>
      .
      .
  </LinkSet>
</eLinkResult>
```

3.3.2 Python

Η Python αποτελεί μια ευρέως διαδεδομένη γλώσσα προγραμματισμού υψηλού επιπέδου. Έχει σχεδιαστεί έτσι ώστε να παρέχει αναγνωσιμότητα και ευκολία στη χρήση της. Διακρίνεται για την ταχύτητα εκμάθησής της και για το μεγάλο πλήθος εύχρηστων βιβλιοθηκών που διαθέτει. Ένα ιδιαίτερο χαρακτηριστικό της γλώσσας είναι η χρήση κενών διαστημάτων (whitespace) για τον διαχωρισμό των συντακτικών δομών του προγράμματος, σε αντίθεση με την πρακτική σε άλλες γλώσσες όπου για τον ίδιο σκοπό χρησιμοποιούνται ειδικά σύμβολα (π.χ. αγκύλες). Αυτό κάνει ένα πρόγραμμα γραμμένο σε Python να μοιάζει περισσότερο με φυσικό κείμενο παρά με κώδικα γλώσσας προγραμματισμού.



Εικόνα 3.1: Η Python διαθέτει μεγάλο πλήθος βιβλιοθηκών....

Η Python αναπτύσσεται ως ανοιχτό λογισμικό (open source) και η διαχείρισή της γίνεται από τον μη κερδοσκοπικό οργανισμό Python Software Foundation από τον οποίο και διανέμεται.

Χαρακτηριστικά

Αποτελεί γλώσσα γενικού σκοπού και μπορεί να υποστηρίξει πολλά μοντέλα προγραμματισμού όπως τον αντικειμενοστραφή, τον δηλωτικό και τον διαδικαστικό προγραμματισμό. Επίσης η γραμματική της δεν είναι αυστηρή, συγκριτικά με αυτή άλλων γλωσσών προγραμματισμού.

Η Python διαθέτει αυτόματη διαχείριση μνήμης κάτι οποίο εξασφαλίζεται από έναν μετρητή αναφορών προς τις μεταβλητές και τα αντικείμενα και από έναν garbage collector. Ένα ακόμη χαρακτηριστικό της Python αποτελεί η δυναμική δέσμευση των ονομάτων των μεταβλητών και των συναρτήσεων (late binding) η οποία πραγματοποιείται κατά την εκτέλεση του αντίστοιχου προγράμματος.

Κάθε υλοποίηση της Python συνοδεύεται και από έναν διερμηνευτή, κάτι το οποίο σημαίνει ότι η μετάφραση ενός προγράμματος γίνεται σε χρόνο εκτέλεσης. Όταν ο κώδικας γίνεται import (εντολή με την οποία εισάγονται βιβλιοθήκες γραμμένες στην Python) παράγεται κώδικας σε συμβολική γλώσσα για μία εικονική μηχανή, ο κώδικας αυτός αποθηκεύεται στον δίσκο και δεν ξαναμεταφράζεται, παρά όταν ξαναγίνει import. Οι διερμηνευτές της Python μπορούν να εγκατασταθούν σε πολλά λειτουργικά συστήματα. Είναι σχεδιασμένη ώστε τόσο αυτή όσο και οι διερμηνευτές της να είναι εύκολα επεκτάσιμοι ώστε να προσαρμόζεται κάθε φορά στις απαιτήσεις.

Στα πλάισια της διπλωματικής χρησιμοποιήθηκε η διανομή 3.3 της Python καθώς και οι παρακάτω βιβλιοθήκες της:

Python NetworkX

Το πακέτο της Python NetworkX χρησιμοποιείται για τη δημιουργία, επεξεργασία και μελέτη της δομής και της λειτουργίας σύνθετων δικτύων. Κάποια από τα χαρακτηριστικά του είναι τα εξής:

- Διαθέτει δομές δεδομένων όπως γράφους, γράφους διπλής κατεύθυνσης, καθώς και γράφους με πολλαπλές ακμές μεταξύ δύο κόμβων.
- Υλοποιεί πολλούς γνωστούς αλγόριθμους της θεωρίας των γράφων.
- Διαθέτει generators κλασσικών ή τυχαίων γράφων καθώς και σύνθετων δικτύων.
- Δίνει τη δυνατότητα ο τύπος των κόμβων να είναι οποιοσδήποτε ακόμα και κείμενο, εικόνες ή XML εγγραφές
- Οι ακμές μπορούν να περιέχουν δεδομένα όπως βάρη.
- Επιπρόσθετα θετικά χαρακτηριστικά κληρονομημένα από την Python, είναι η ευκολία στην επέκταση και στη εκμάθησή του καθώς και η ανεξαρτησία πλατφόρμας.

Στα πλαίσια της διπλωματικής χρησιμοποιήσαμε την έκδοση 1.8.1 της βιβλιοθήκης αυτής.

Χρησιμοποιήσαμε επίσης τη δομή δεδομένων της που υλοποιεί γράφους διπλής κατεύθυνσης (DiGraph()) καθώς και διάφορες μεθόδους προσπέλασης και επεξεργασίας αυτών των γράφων, όπως προσθήκη κόμβων από λίστα (add_nodes_from) και προσθήκη ακμών (add_edge).

Ο κύριος λόγος χρήσης αυτής της βιβλιοθήκης ήταν η ενσωματωμένη μέθοδός της που υλοποιεί τον αλγόριθμο Pagerank. Τα ορίσματα της μεθόδου Pagerank είναι τα εξής:

G: Ένας NetworkX γράφος.

alpha: Η παράμετρος απόσβεσης του PageRank, τύπου float. Είναι προαιρετικό όρισμα με default τιμή 0.85.

personalization: Το όρισμα αυτό είναι επίσης προαιρετικό και αποτελεί ένα διάνυσμα τύπου dictionary με keys τους κόμβους του γράφου και values μη μηδενικές τιμές για την παραμετροποίηση της εισόδου. Οι προεπιλεγμένες τιμές ακολουθούν ομοιόμορφη κατανομή.

max_iter: Ο μέγιστος αριθμός των επαναλήψεων του αλγορίθμου. Είναι προαιρετικό όρισμα τύπου integer.

tol: Αποτελεί προαιρετικό όρισμα για τη μέγιστη επιτρεπτή απόκλιση του αλγόριθμου κατά τον τερματισμό του.

nstart: Αποτελεί επίσης προαιρετικό όρισμα τύπου dictionary με αρχικές τιμές εκκίνησης του αλγόριθμου για κάθε κόμβο του γράφου.

weight: Βάρη των ακμών του γράφου (προαιρετικό).

dangling: Σε αυτό το όρισμα δηλώνεται αν υπάρχουν κόμβοι χωρίς εξερχόμενες ακμές. Είναι τύπου dictionary με keys τους κόμβους του γράφου και values 0 αν ο κόμβος

έχει εξερχόμενες ακμές ή 1 αν ο κόμβος δεν έχει εξερχόμενες ακμές. Σε περίπτωση που αυτό το όρισμα δεν χρησιμοποιείται ο αλγόριθμός υπολογίζει μόνος του τους κόμβους χωρίς ακμές.

Η μέθοδος επιστρέφει ένα dictionary με keys τους κόμβους και values τις τιμές του PageRank για κάθε κόμβο.

MySQL Connector

Το πακέτο MySQL Connector/Python δίνει τη δυνατότητα τα προγράμματα γραμμένα σε Python να έχουν πρόσβαση σε MySQL βάσεις δεδομένων. Ο κωδικάς του είναι γραμμένος σε Python με αποτέλεσμα να μην έχει εξαρτήσεις από άλλες βιβλιοθήκες παρά μόνο από τη standard βιβλιοθήκη της Python.

Η βιβλιοθήκη MySQL Connector/Python παρέχει υποστήριξη για τις παρακάτω λειτουργίες:

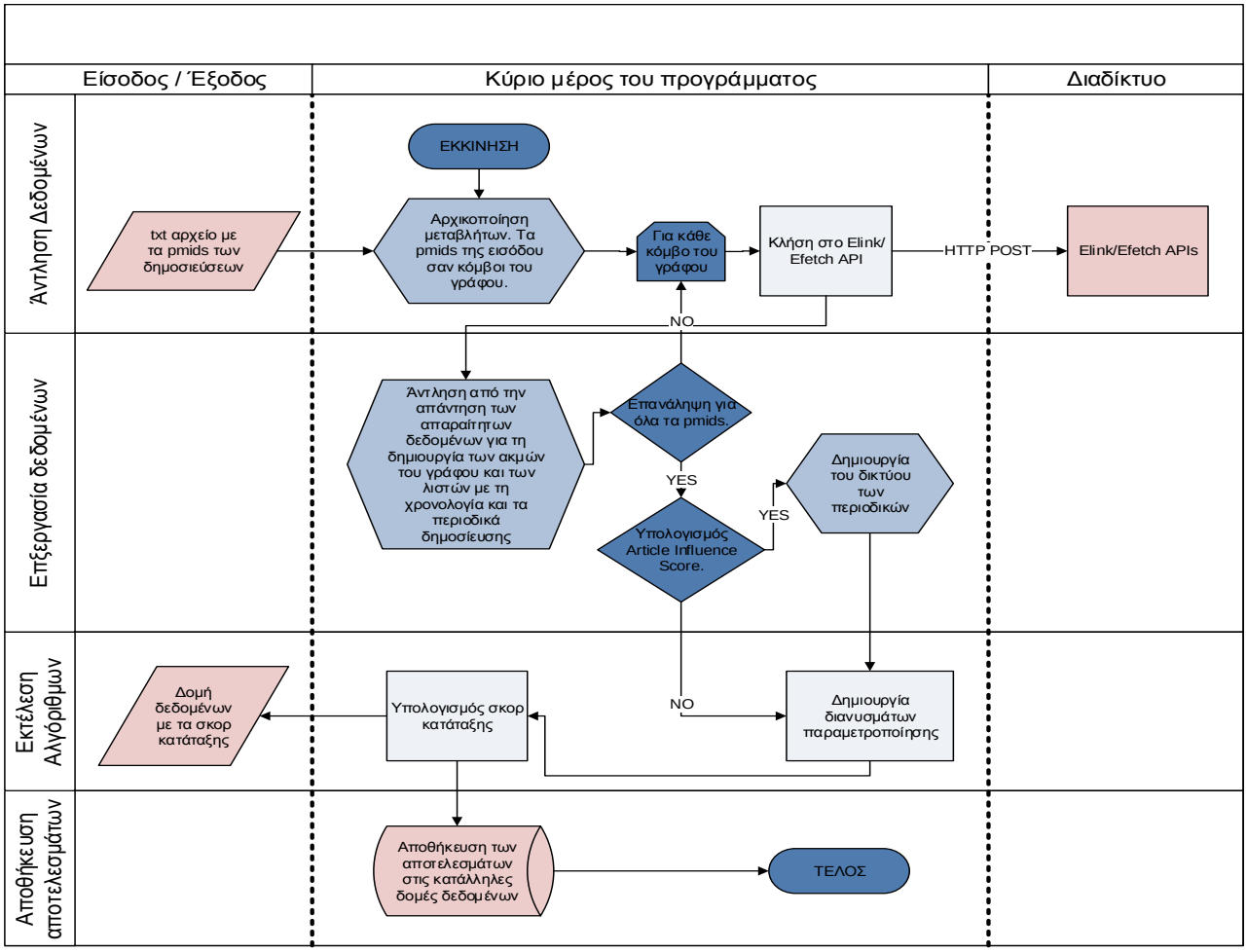
- MYSQL server από την έκδοση 5.7 και μετά.
- Μετατροπή όλων των τύπων των μεταβλητών της Python στους αντίστοιχους της MySQL και αντίστροφα. Υπάρχει ρύθμιση που επιτρέπει η παραπάνω διαδικασία να γίνεται αυτόματα για ευκολία ή το αντίθετο σε περίπτωση που αναζητάμε τη βέλτιστη απόδοση.
- Επεκτάσεις της MySql που αφορούν στο συντακτικό της SQL.
- Πρωτόκολλο συμπίεσης που χρησιμεύει στη συμπίεση των δεδομένων κατά τη μεταφορά τους μεταξύ πελάτη και εξυπηρετητή.
- Σύνδεση μεταξύ πελάτη-εξυπηρετητή με βάση το TCP/IP πρωτόκολλο και την τεχνολογία SSL.
- Δεν απαιτείται η επιπλέον χρήση κάποιας βιβλιοθήκης στο σύστημα του πελάτη.

Χρησιμοποιήθηκε η έκδοση 2.0.3 της βιβλιοθήκης αυτής, σε υλοποίηση που περιγράφεται στο Κεφάλαιο 5.

3.4 Περιγραφή υλοποίησης

Δημιουργήσαμε ένα Python script με συναρτήσεις τους αλγόριθμους των ενοτήτων 3.1.1 έως 3.1.7. Κάθε συνάρτηση παίρνει σαν είσοδο το όνομα ενός TXT αρχείου με τα pmids των δημοσιεύσεων (αναγνωριστικά των δημοσιεύσεων στη βάση PubMed), που είναι καταχωρημένες στη βάση και επιστρέφει μία δομή δεδομένων dictionary της Python, με κλειδιά τα pmids των δημοσιεύσεων και τιμές τα αντίστοιχα σκορ της μετρικής κατάταξης που χρησιμοποιήθηκε. Όλες οι συναρτήσεις αν και υλοποιούν διαφορετικούς αλγόριθμους ακολουθούν μία κοινή διαδικασία. Αρχικά συλλέγονται τα δεδομένα που απαιτούνται για την είσοδο των αλγόριθμων, τα οποία οι συναρτήσεις επεξεργάζονται και τα μετατρέπουν στην κατάλληλη μορφή. Στη συνέχεια υπολογίζονται τα σκορ κατάταξης για όλες τις δημοσιεύσεις της εισόδου και τέλος τα σκορ αυτά καθώς και οι δομές δεδομένων της εισόδου του αλγορίθμου

που κατασκευάστηκαν, αποθηκεύονται τοπικά. Στο Σχήμα 3.1 παρουσιάζεται το διάγραμμα ροής με τις λειτουργίες του προγράμματος.



Σχήμα 3.1: Διάγραμμα ροής του υλοποιημένου προγράμματος

3.4.1 Συλλογή δεδομένων

Έχοντας σαν δεδομένο τη λίστα των αναγνωριστικών των άρθρων στη βάση Pubmed (pmids) που είναι καταχωρημένα και στην εφαρμογή DIANA mirPub, ζητούμενο ήταν να συλλέξουμε το σύνολο των αναφορών, τη χρονολογία και το περιοδικό δημοσίευσης των άρθρων αυτών. Γι' αυτό το λόγο χρησιμοποιήσαμε τα E-Fetch και E-Link utilities της μηχανής αναζήτησης Entrez (3.3.1). Συγκεκριμένα, με τη χρήση του πρώτου προγράμματος, στέλνοντας μία HTTP POST αίτηση με headers τα ορίσματα του προγράμματος που περιγράψαμε στην προηγούμενη ενότητα (db=pubmed, retmode=xml και id το pmid του κάθε άρθρου) παίρνουμε σαν αποτέλεσμα να XML αρχείο που περιέχει όλες τις ζητούμενες πληροφορίες όπως αυτό της Ενότητας 3.3.1.1.

Τα δεδομένα που μας ενδιαφέρουν από το παραπάνω αρχείο είναι αυτά που βρίσκονται στα πεδία DateCreated/Year (έτος δημοσίευσης), Journal/Title (τίτλος περιοδικού δημοσίευσης)

και το σύνολο των πεδίων CommentsCorrections/PMID (τα αναγνωριστικά pmids των παραπομπών του άρθρου). Από το σύνολο των CommentsCorrections πεδίων, μας ενδιαφέρουν μόνο όσα έχουν σαν attribute RefType="Cites" ή RefType="CommentOn" διότι μόνο σε αυτή την περίπτωση τα αντίστοιχα άρθρα αποτελούν αναφορές.

Επειδή παρατηρήσαμε ότι με το eutil E-Link παίρνουμε διαφορετικό σύνολο από αναφορές, για να εμπλουτίσουμε τις ακμές του γράφου που πρόκειται να δημιουργήσουμε χρησιμοποιήσαμε και αυτό. Πραγματοποιήσαμε λοιπόν και σε αυτή την περίπτωση μια HTTP POST αίτηση για κάθε άρθρο. Μέρος της απάντησης του E-Link API παρατίθεται στην Ενότητα 3.3.1.2.

Όπως φαίνεται στο συγκεκριμένο αρχείο, για κάθε άρθρο επιστρέφονται πολλά σύνολα από συνδέσμους σχετικούς με αυτό. Το σύνολο που περιέχει τις παραπομπές του άρθρου, που βρίσκονται τόσο στην Pubmed βάση όσο και στην PMC και ώστε να μπορούμε να συλλέξουμε τις αντίστοιχες πληροφορίες και γι' αυτές, είναι αυτό με LinkName pubmed_pubmed_refs.

3.4.2 Επεξεργασία δεδομένων

Αφού συλλέξαμε τα απαιτούμενα XML αρχεία για κάθε pmid της εισόδου, στη συνέχεια τα προσπελάσαμε με τη χρήση της ενσωματωμένης βιβλιοθήκης της Python xml.etree.ElementTree. Με τα δεδομένα που εξάγαμε για κάθε άρθρο, δημιουργήσαμε έναν γράφο της NetworkX βιβλιοθήκης του οποίου οι κόμβοι αποτελούνται από τα pmids των άρθρων και οι ακμές του από τις αναφορές του κάθε άρθρου. Αξίζει να σημειώσουμε ότι τα XML αρχεία περιείχαν μεγάλο πλήθος αναφορών για κάθε άρθρο. Εμείς κρατήσαμε μόνο αυτές που απευθύνονται σε δημοσιεύσεις που υπάρχουν στη βάση της εφαρμογής Mirpub και επομένως στη λίστα εισόδου του προγράμματος. Όσον αφορά τις υπόλοιπες πληροφορίες που συλλέξαμε για κάθε άρθρο, δημιουργήσαμε δύο δομές δεδομένων τύπου dictionary της Python με κλειδιά τα pmids των άρθρων και τιμές τα έτη και τα περιοδικά δημοσίευσης αντίστοιχα.

Για τον υπολογισμό του Article Influence score των περιοδικών είναι απαραίτητη η δημιουργία ενός επιπλέον γράφου αντίστοιχου με αυτόν των δημοσιεύσεων. Ο γράφος έχει σαν κόμβους τα περιοδικά στα οποία είναι δημοσιευμένα τα άρθρα της εισόδου και ακμές τις αναφορές των άρθρων ενός περιοδικού προς τα άρθρα των άλλων περιοδικών. Οι πολλαπλές ακμές που μπορεί να υπάρχουν ανάμεσα σε δύο περιοδικά εκφράζονται σαν βάρος της αντίστοιχης ακμής το οποίο έχει σαν τιμή το ποσοστό των αναφορών των άρθρων του πρώτου περιοδικού προς το δεύτερο, προς το σύνολο των αναφορών των άρθρων του πρώτου περιοδικού.

Απαιτείται επίσης και μια δομή δεδομένων στην οποία το όνομα κάθε περιοδικού αντιστοιχίζεται με το ποσοστό των άρθρων που είναι δημοσιευμένα στο συγκεκριμένο περιοδικό, προς το σύνολο των άρθρων.

3.4.3 Υλοποίηση αλγόριθμων

Δημιουργία παραμετροποίησης που προωθεί τις νέες δημοσιεύσεις

Υλοποιήθηκε η συνάρτηση που λαμβάνει σαν είσοδο το διάνυσμα με τις χρονολογίες δημοσίευσης των άρθρων και επιστρέφει το παρακάτω διάνυσμα που περιγράφεται στην

Ενότητα 3.1:

$$\mathbf{w}[i] = \frac{e^{(\mathbf{y}[i]-c)/b}}{b}$$

Συνάρτηση υπολογισμού του impact factor των περιοδικών

Υλοποιήθηκε η συνάρτηση που λαμβάνει σαν είσοδο τα διανύσματα με τις χρονολογίες και τα περιοδικά δημοσίευσης των άρθρων, τον γράφο που δημιουργήθηκε, το διάστημα στο οποίο θέλουμε να υπολογιστεί η μετρική (2 ή 5 έτη) καθώς και το τρέχον έτος (0 αν θέλουμε να είναι μεταβαλλόμενο), υπολογίζει το impact factor όπως περιγράφεται από τους τύπους 3.3,3.4 και 3.5, ανάλογα με την είσοδο και επιστρέφει ένα διάνυσμα κανονικοποιημένο στο διάστημα (1,9) (θέλαμε να αποφύγουμε την παραγωγή μηδενικών σκορ για περιοδικά των οποίων τα άρθρα δεν αποτελούν παραπομπές), με την τιμή της μετρικής impact factor του περιοδικού για κάθε άρθρο.

Συνάρτηση υπολογισμού Article Influence score

Υλοποιήθηκε η συνάρτηση που λαμβάνει τον γράφο των περιοδικών, και το διάνυσμα με το ποσοστό του πλήθους των άρθρων που είναι δημοσιευμένα σε κάθε περιοδικό, υπολογίζει το eigenfactor των περιοδικών εφαρμόζοντας ουσιαστικά τον αλγόριθμο PageRank στον γράφο των περιοδικών με συντελεστή απόσβεσης $a = 0.85$, λαμβάνοντας υπόψη και τα βάρη των ακμών και στη συνέχεια παράγει το Article Influence score. Τέλος επιστρέφει ένα διάνυσμα κανονικοποιημένο στο διάστημα (1,9), με την τιμή της μετρικής του περιοδικού δημοσίευσης κάθε άρθρου.

Συνάρτηση υπολογισμού PageRank

Χρησιμοποιήθηκε η συνάρτηση PageRank της βιβλιοθήκης NetworkX και εισάγοντας σε αυτήν το κατάλληλο διάνυσμα παραμετροποίησης (personalization vector) υλοποιήθηκαν οι αλγόριθμοι **AdvRecPubs**, **RankWithIF**, **RankWithAI**, **AdvRecPubs-RankWithIF**, **FutureRankLike** και **Weighted Citations in PageRank**.

Δημιουργία αλγόριθμου κατάταξης σε γράφο με σταθμισμένες ακμές

Δημιουργήθηκε συνάρτηση που λαμβάνει σαν είσοδο τον γράφο των άρθρων, καθώς και τα διανύσματα της χρονολογίας και των περιοδικών δημοσίευσής τους, εισάγει βάρη στις ακμές του γράφου αυτού σύμφωνα με τον τύπο:

$$weight = e^{-0.0117(y[a]-y[b])}$$

όπου a, b ο αρχικός και τελικός κόμβος της ακμής αυτής και $\mathbf{y}$ το διάνυσμα με τις χρονολογίες δημοσίευσης των περιοδικών.

Στη συνέχεια παράγει την απαιτούμενη είσοδο (τον γράφο των περιοδικών και το ποσοστό των άρθρων που είναι δημοσιευμένα σε κάθε άρθρο) και καλεί την συνάρτηση που υπολογίζει το Article Influence score. Τέλος, υπολογίζει το σκορ με βάση το Σχήμα 2.10.

Συνάρτηση υπολογισμού FutureRankLike

Τροποποιήσαμε τη συνάρτηση PageRank της βιβλιοθήκης NetworkX ώστε να δέχεται σαν είσοδο και τον συντελεστή απόσβεσης b, καθώς κι έναν επιπλέον όρο, ο οποίος στη συγκεκριμένη περίπτωση είναι η μετρική δημοτικότητας-κύρους των περιοδικών.

3.4.4 Αποθήκευση δεδομένων εισόδου και αποτελεσμάτων

Ο γράφος και τα διανύσματα που δημιουργήθηκαν κατά την επεξεργασία των αρχείων από τα Entrez e-utilities αποθηκεύτηκαν τοπικά σε δυαδική μορφή με τη χρήση της ενσωματωμένης βιβλιοθήκης της Python, pickle. Με αυτό τον τρόπο, τα δεδομένα μπορούν να ανακτηθούν εύκολα στην επιθυμητή μορφή και να επαναχρησιμοποιηθούν. Αυτό μας έδωσε τη δυνατότητα να παραλείπουμε τα βήματα της συλλογής δεδομένων και δημιουργίας των κατάλληλων δομών δεδομένων όταν τρέχουμε τους αλγόριθμους για κοινό σύνολο δημοσιεύσεων.

Τέλος αποθηκεύσαμε και τα αποτελέσματα του κάθε αλγόριθμου τόσο σε δυαδική μορφή όσο και σε μορφή TXT αρχείου, κάτι το οποίο είναι χρήσιμο για τη σύγκριση και αξιολόγηση των αποτελεσμάτων που πραγματοποιείται στο Κεφάλαιο 4, χωρίς την ανάγκη επανεκτέλεσης των προγραμμάτων κάθε φορά που θέλουμε να εξάγουμε διαφορετικό συμπέρασμα.

Κεφάλαιο 4

Αξιολόγηση υλοποιημένων αλγορίθμων

Πέρα από την υλοποίηση των αλγορίθμων κατάταξης δημοσιεύσεων, σημαντικό μέρος αυτής της εργασίας ήταν και η αξιολόγησή τους. Η αξιολόγηση γίνεται με σκοπό να αποφανθούμε αν οι οι αλγόριθμοι αυτοί, κατατάσσουν πιο αξιόπιστα από τον Pagerank τις δημοσιεύσεις με βάση τη σημαντικότητά τους, δεδομένου ότι ευνοούν τις νεότερες από αυτές. Τέλος, με βάση την αξιολόγηση, επιλέγουμε να ενσωματώσουμε στην εφαρμογή αυτούς που επιστρέφουν τις καλύτερες δυνατές κατατάξεις. Η αξιολόγηση πραγματοποιείται με τη σύγκριση της κατάταξης - εξόδου κάθε αλγόριθμου με κάποια «ιδανική κατάταξη». Απαιτείται λοιπόν η δημιουργία μιας μεθόδου αξιολόγησης, η οποία πρόκειται να ορίζει την «ιδανική κατάταξη» και να συγκρίνει τα αποτελέσματα των αλγόριθμων με αυτήν. Στο κεφάλαιο αυτό, περιγράφεται η μέθοδος αξιολόγησης που εφαρμόστηκε μαζί με της μετρικές που χρησιμοποιεί. Τέλος, παρατίθενται τα αποτελέσματα και τα συμπεράσματα που προκύπτουν από αυτή τη διαδικασία αξιολόγησης.

4.1 Μετρικές σύγκρισης ταξινομημένων λιστών

Στην ενότητα αυτή περιγράφονται οι μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση των αποτελεσμάτων των υλοποιημένων αλγόριθμων: ο συντελεστής συσχέτισης Spearman που χρησιμοποιείται στην ανάλυση της ομοιότητας δύο μεταβλητών και οι μετρικές «ανάκληση» και «ακρίβεια» με κύρια εφαρμογή στην ανάκτηση πληροφορίας.

4.1.1 Ο συντελεστής συσχέτισης Spearman

Ο συντελεστής συσχέτισης Spearman ή Spearman's ρ, όπως συναντάται στη βιβλιογραφία [26][14], αποτελεί μετρική της στατιστικής εξάρτησης δύο μεταβλητών και αξιολογεί τη δυνατότητα περιγραφής της μεταξύ τους σχέσης με μία μονότονη συνάρτηση. Χρησιμοποιείται κυρίως για τη σύγκριση μεταβλητών κατάταξης στοιχείων και βασίζεται στην Ευκλείδεια α-

πόστασή τους. Ο τύπος της Ευκλείδειας απόστασης δίνεται από τον τύπο:

$$\rho(\sigma_1, \sigma_2) = \sqrt{\sum_{i=1}^{n} |\sigma_1(i) - \sigma_2(i)|^2} \tag{4.1}$$

όπου $\sigma_1(i)$ και $\sigma_2(i)$ η κατάταξη του στοιχείου i με βάση την κάθε μεταβλητή.

Ο συντελεστής Spearman διαμορφώνεται χρησιμοποιώντας σαν βάση την απόσταση αυτή και πραγματοποιώντας την ακόλουθη κανονικοποίηση στο διάστημα $[-1, 1]$:

$$\rho(\sigma_1, \sigma_2) = 1 - \frac{6 \sum_{i=1}^{n} |\sigma_1(i) - \sigma_2(i)|^2}{n(n^2 - 1)} \tag{4.2}$$

Στα πλαίσια της εργασίας, ο συντελεστής συσχέτισης Spearman χρησιμοποιείται για τη σύγκριση των κατατάξεων των δημοσιεύσεων που προκύπτουν από την ταξινόμησή τους με βάση τα σκορ των υλοποιημένων αλγόριθμων. Σε μία κατάταξη η κάθε δημοσίευση καταλαμβάνει και μία συγκεκριμένη θέση. Ακόμα και δημοσιεύσεις για τις οποίες ο υλοποιημένος αλγόριθμος επιστρέφει το ίδιο σκορ, καταλαμβάνουν διαφορετική θέση που καθορίζεται με τυχαίο τρόπο αφού στον χρήστη εμφανίζονται με τη μορφή λίστας. Συνεπώς, μπορεί να δημιουργηθούν λανθασμένα μεγάλες αποστάσεις σ_1 και σ_2 για κάποιες δημοσιεύσεις. Στην περίπτωση λοιπόν που οι δημοσιεύσεις έχουν το ίδιο σκορ, θεωρούμε ότι έχουν και κοινή θέση στην κατάταξη, που ισούται με το άθροισμα των θέσεων των δημοσιεύσεων με κοινό σκορ προς το πλήθος τους. Αν για παράδειγμα οι δημοσιεύσεις που βρίσκονται στις θέσεις 10 έως 12 μιας κατάταξης έχουν κοινό σκορ θεωρούμε ότι όλες βρίσκονται στη θέση $\frac{10+11+12}{3} = 11$.

Δεν είναι πάντα απαραίτητη ο σύγκριση του συνόλου των κατατάξεων παρά μόνο των top-k αποτελεσμάτων τους. Στα πειράματα που διεξήχθησαν στα πλαίσια της διπλωματικής, χρησιμοποιήσαμε την επέκταση του υπολογισμού του Spearman που έχει προταθεί στην εργασία [26] και εφαρμόζεται σε top-k αποτελέσματα κατατάξεων. Σε αυτήν την περίπτωση, οι κατατάξεις που τίθενται προς σύγκριση, οι οποίες αποτελούν υποσύνολο των γενικών κατατάξεων έχουν και μη κοινά στοιχεία μεταξύ τους. Όταν ένα στοιχείο περιλαμβάνεται μόνο σε μία εκ των δύο κατατάξεων θεωρούμε ότι η θέση του στην άλλη είναι ίση με $k+1$, όταν συγκρίνουμε τα top-k αποτελέσματα των κατατάξεων.

Ενώ ο τύπος 4.1 για τη μετρική Spearman εξακολουθεί να ισχύει, η κανονικοποίση δεν μπορεί να είναι η ίδια μιας και η μέγιστη τιμή της μετρικής διαφέρει με βάση τα παραπάνω. Η μέγιστη τιμή σε αυτήν την περίπτωση ισούται με:

$$\rho(\sigma_1, \sigma_2) = \sqrt{\frac{k(k+1)(2k+1)}{3}} \tag{4.3}$$

Επομένως στην περίπτωση της εφαρμογής του τύπου Spearman σε top-k αποτελέσματα, αυτός διαμορφώνεται ως:

$$\rho(\sigma_1, \sigma_2) = 1 - \frac{6 \sum_{i=1}^{n} |\sigma_1(i) - \sigma_2(i)|^2}{k(k+1)(2k+1)} \tag{4.4}$$

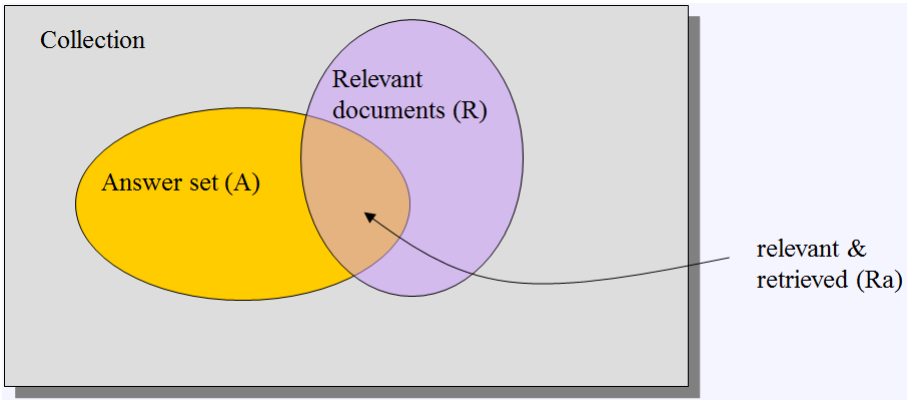
4.1.2 Καμπύλη Ακρίβειας-Ανάκλησης

Μία ακόμη μετρική που χρησιμοποιήθηκε για την αξιολόγηση των αποτελεσμάτων είναι η καμπύλη precision - recall (καμπύλη Ακρίβειας - Ανάκλησης). Η καμπύλη αυτή κανονικά χρησιμοποιείται για την αποτίμηση της αποτελεσματικότητας ενός συστήματος ως προς την ανάκτηση πληροφορίας με βάση τα ερωτήματα που εισάγει ο χρήστης. Παρ' όλα αυτά, προσαρμόσαμε τα παραπάνω μεγέθη ώστε να αξιολογήσουμε την κατάταξη που πραγματοποιήσαμε στα κείμενα του συστήματός μας.

Οι συνήθεις ορισμοί της Ακρίβειας (precision) και της Ανάκλησης (recall) βασίζονται στις ακόλουθες παραδοχές:

Έστω το σύνολο των κειμένων του εξεταζόμενου συστήματος. Όταν πραγματοποιείται αναζήτηση στο σύστημα αυτό με βάση κάποιο ερώτημα, τα κείμενα αυτά χαρακτηρίζονται ως:

- *«Σχετικά» ή «μη σχετικά» με το ερώτημα κείμενα.*
- *«Επιλεγμένα» ή «μη επιλεγμένα» από το σύστημα κείμενα.*
- *Ως «γενικότητα» ορίζεται επίσης το ποσοστό επί του συνόλου των κειμένων τα οποία θεωρούνται σχετικά με το ερώτημα.*



Σχήμα 4.1: Τα υποσύνολα των κειμένων που δημιουργούνται με την εφαρμογή ενός ερωτήματος.

Ορισμός 4.3. *Για οποιοδήποτε σύνολο επιλεγμένων κειμένων με βάση κάποιο ερώτημα, ως ανάκληση ορίζεται το ποσοστό των σχετικών με το ερώτημα κειμένων που επιστρέφονται από το σύστημα ως αποτέλεσμα της αναζήτησης, επί του συνόλου των σχετικών με το ερώτημα κειμένων του συστήματος.*

$$Recall = \frac{Ra}{R} \tag{4.5}$$

Μπορεί να επιτευχθεί 100% ανάκληση όταν ανακτηθεί το σύνολο των κειμένων του συστήματος, όπου εμπεριέχεται και το σύνολο των σχετικών με το ερώτημα κειμένων, αλλά αυτό ακυρώνει το σκοπό του συστήματος ανάκτησης. Συμπερασματικά αυτό που είναι επιθυμητό είναι μεγάλο ποσοστό ανάκλησης κατά την εκκίνηση της διαδικασίας της ανάκτησης.

Ορισμός 4.4. *Για οποιοδήποτε σύνολο επιλεγμένων κειμένων με βάση κάποιο ερώτημα, ως ακρίβεια ορίζεται το ποσοστό των σχετικών με το ερώτημα κειμένων που επιστρέφονται από το σύστημα ως αποτέλεσμα της αναζήτησης, επί του συνόλου των κειμένων που επιστρέφονται.*

$$Precision = \frac{Ra}{A} \tag{4.6}$$

Στόχος λοιπόν ενός συστήματος ανάκτησης είναι να μεγιστοποιεί και τα δύο αυτά μεγέθη. Έστω λοιπόν ότι έχουμε μια μηχανή αναζήτησης κειμένων των οποίων το σύνολο ισούται με 1000 και πραγματοποιούμε ένα ερώτημα για το οποίο τα 100 κείμενα θεωρούνται σχετικά. Παρακάτω θα περιγράψουμε πώς μεταβάλλεται η Ακρίβεια και η Ανάκληση σε σχέση με το πλήθος των κειμένων που ανακτώνται για 3 περιπτώσεις: την ιδανική, τη χείριστη δυνατή και την τυχαία ανάκτηση.

Ιδανική ανάκτηση Σε αυτή την περίπτωση, όλα τα σχετικά με το ερώτημα κείμενα ανακτώνται πριν από το πρώτο μη σχετικό.

Η ανάκληση αυξάνεται με σταθερό ρυθμό καθώς αυξάνεται το σύνολο των κειμένων που ανακτώνται και συγκεκριμένα ίσο με το αντίστροφο της γενικότητας όπως έχει οριστεί παραπάνω (1000/100), μέχρις ότου η τιμή της να γίνει ίση με τη μονάδα όπου όλα τα σχετικά κείμενα (100) έχουν ανακτηθεί. Η τιμή της ανάκλησης από εκεί και πέρα παραμένει σταθερή αφού, σύμφωνα με τον τύπο της, τα επιπλέον κείμενα δεν την επηρεάζουν.

Η ακρίβεια ξεκινά με σταθερή τιμή, ίση με τη μονάδα, μέχρις ότου όλα τα σχετικά κείμενα ανακτηθούν (100). Από εκεί και πέρα η καμπύλη της ακρίβειας παίρνει τη μορφή της υπερβολικής συνάρτησης $\frac{1}{x}$ δεδομένου ότι ο ο αριθμητής του κλάσματος που ορίζει την ακρίβεια παραμένει σταθερός και ίσος με 100 ενώ αυξάνεται ο παρονομαστής, το πλήθος δηλαδή των κειμένων που ανακτώνται. Όταν ανακτηθούν και τα 1000 κείμενα η τιμή της Ακρίβειας ισούται με αυτή της Γενικότητας.

Χείριστη δυνατή ανάκτηση Σε αυτή την περίπτωση, όλα τα μη σχετικά με το ερώτημα κείμενα ανακτώνται πριν από το πρώτο σχετικό.

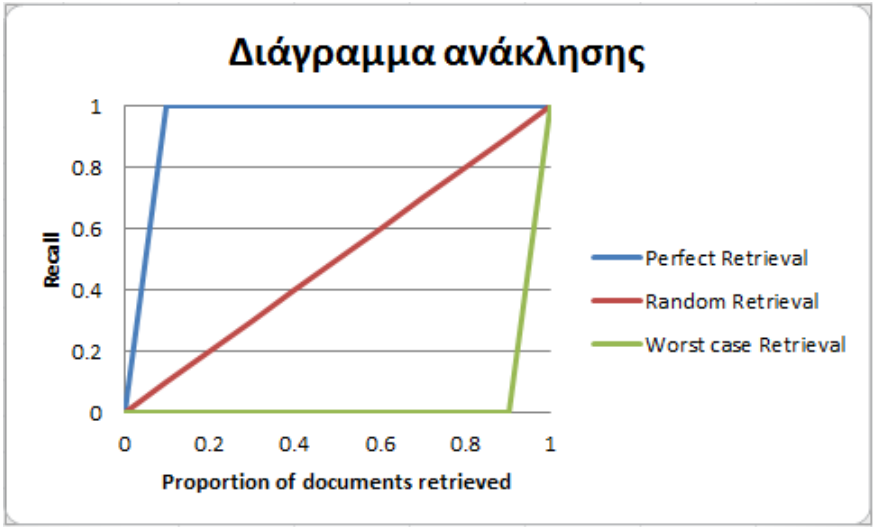
Η ανάκληση έχει μηδενική τιμή όσο ανακτώνται μόνο μη σχετικά κείμενα. Όταν ο αριθμός των κειμένων που επιστρέφονται ξεπερνά τα 900, αυτά δεν μπορούν παρά να είναι σχετικά, με αποτέλεσμα η τιμή της ανάκλησης να αυξάνεται με σταθερό ρυθμό και ίσο με το αντίστροφο της γενικότητας (1000/100) μέχρι να ανακληθεί και το τελευταίο κείμενο όπου ισούται με τη μονάδα.

Η ακρίβεια επίσης ξεκινά με μηδενική τιμή, η οποία παραμένει σταθερή για τα πρώτα 900 κείμενα. Στη συνέχεια η γραφική απεικόνισή της ακολουθεί τη συνάρτηση $f(x) = \frac{x-900}{x}$ μέχρι να ανακτηθούν και τα 100 σχετικά κείμενα και ο αριθμητής του τύπου της να ισούται με 100 ενώ ο παρονομαστής με 1000.

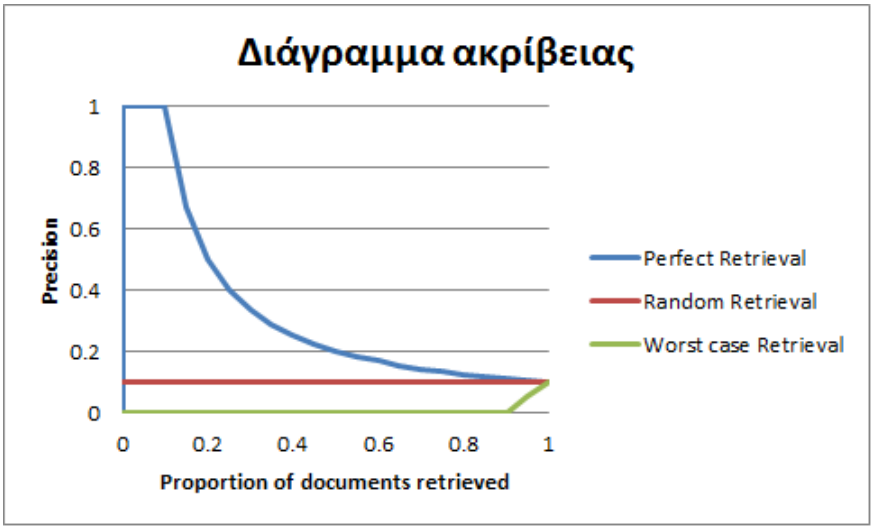
Τυχαία ανάκτηση Σε αυτή την περίπτωση, η πιθανότητα να ανακτηθεί ένα σχετικό με το ερώτημα κείμενο είναι ίδια για όλη τη διάρκεια της διαδικασίας της ανάκτησης. Συγκεκριμένα είναι ίση με $\frac{100}{1000} = 0.1$.

Όταν λοιπόν έχουν ανακτηθεί x κείμενα, σχετικά από αυτά είναι τα $0.1x$ επομένως κάθε στιγμή ο τύπος της ανάκλησης παίρνει τη μορφή $Recall = \frac{0.1x}{100}$.

Αντίστοιχα $Precision = \frac{0.1x}{x} = 0.1$



Σχήμα 4.2: Η Ανάκληση με βάση διαφορετικούς τρόπους ανάκτησης.



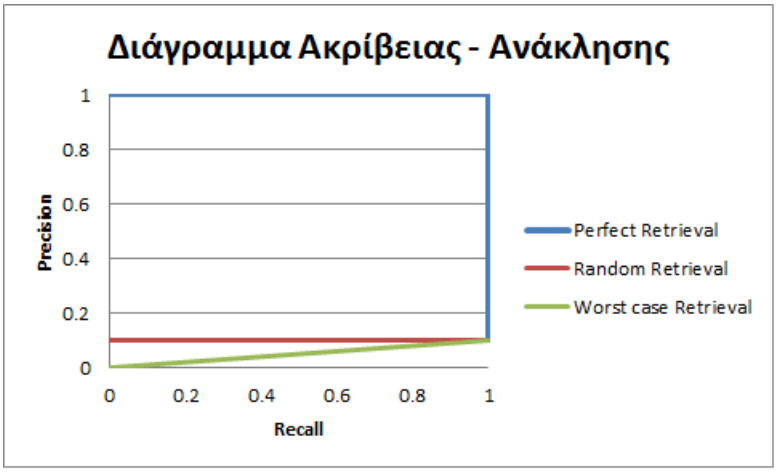
Σχήμα 4.3: Η Ακρίβεια με βάση διαφορετικούς τρόπους ανάκτησης.

Σε όλες τις περιπτώσεις, οι καμπύλες της ανάκλησης και της ακρίβειας ενός συστήματος δεν μπορούν παρά να περικλείονται από τις αντίστοιχες καμπύλες της τέλειας και χείριστης ανάκτησης. Για να θεωρηθεί ένα σύστημα ανάκτησης αποτελεσματικό πρέπει συμπληρωματικά οι καμπύλες της ακρίβειας και της ανάκλησης να βρίσκονται πάνω από τις αντίστοιχες καμπύλες της τυχαίας ανάκτησης.

Παρατηρούμε επίσης ότι, ο ρυθμός αύξησης της ανάκλησης πρέπει να είναι μεγαλύτερος στην αρχή της διαδικασίας και σταδιακά να μειώνεται. Αντίστοιχα ο ρυθμός μείωσης της ακρίβειας πρέπει να είναι μεγαλύτερος στην αρχή και σταδιακά να μειώνεται. Ο παραπάνω συλλογισμός είναι λογικός αφού θεωρούμε ένα σύστημα αποτελεσματικό όταν όσο το δυνατόν

γρηγορότερα μας επιστρέφει τα σχετικά κείμενα ενώ δεν μας ενδιαφέρει πόσο σχετικά είναι τα κείμενα που μας επιστρέφει στο τέλος.

Εμπειρικά έχουμε καταλήξει στο συμπέρασμα ότι η Ανάκληση και η Ακρίβεια αποτελούν αντιστρόφως ανάλογα μεγέθη [17]. Στο Σχήμα 4.4 παρουσιάζεται η Ακρίβεια έναντι της Ανάκλησης για τις 3 περιπτώσεις που έχουμε αναφέρει. Όπως παρατηρούμε, η αντίστροφη σχέση μεταξύ των δύο αυτών μεγεθών παύει να ισχύει όταν το σύστημα συμπεριφέρεται «χειρότερα» από αυτό που πραγματοποιεί τυχαία ανάκτηση, κάτι το οποίο δεν είναι επιτρεπτό.



Σχήμα 4.4: Ακρίβεια - Ανάκληση με διαφορετικούς τρόπους ανάκτησης.

4.2 Αξιολόγηση αποτελεσμάτων μηχανισμού κατάταξης περιοδικών

Στους αλγόριθμους RankWithIF, AdvRecPubs-RankWithIF και FututreRankLike χρησιμοποιήθηκε η δημοτικότητα των περιοδικών στα οποία είναι δημοσιευμένες οι εργασίες, βασισμένη στις τιμές του Impact Factor των περιοδικών αυτών. Για τον υπολογισμό του Impact Factor, χρησιμοποιήθηκαν μόνο οι δημοσιεύσεις που είναι καταχωρημένες στην εφαρμογή και οι μεταξύ τους παραπομπές. Επομένως, αν και ο υπολογισμός της μετρικής θεωρείται τετριμμένος, κρίνεται απαραίτητη η σύγκριση των σκορ που εξήχθησαν, με τις τιμές του Impact Factor όπως αναρτώνται στο Διαδίκτυο [10] όπου το πλήθος των δεδομένων εισόδου για τον υπολογισμό της μετρικής είναι πολύ μεγαλύτερο και θεωρούμε ότι αποτελούν τις «πραγματικές τιμές» του Impact Factor.

Στους παρακάτω πίνακες παρατίθενται ενδεικτικά τα 10 από το σύνολο των περιοδικών που μας ενδιαφέρουν, με το υψηλότερο Impact Factor, υπολογισμένο για το έτος 2008 (εξηγείται στην Ενότητα 4.4 ο λόγος επιλογής του συγκεκριμένου έτους). Στον Πίνακα 4.1 παρουσιάζονται οι τιμές του Impact Factor που υπολογίστηκαν στα πλαίσια της διπλωματικής ενώ στον 4.2 αυτές που αντλήθηκαν από το Διαδίκτυο.

Κατάταξη	Τίτλος	Impact Factor
1	Cancer cell	14.5
2	Cell metabolism	14
3	Cell	11.4667
4	Journal of clinical oncology : official journal of the American Society of Clinical Oncology	11
5	Molecular cell	10.7273
6	Nature reviews. Cancer	10
7	Nature cell biology	10
8	Nature	9.8667
9	Gastroenterology	9
10	FASEB journal : official publication of the Federation of American Societies for Experimental Biology	9

Πίνακας 4.1: Τα περιοδικά με το υψηλότερο Impact Factor υπολογισμένο με βάση τα δεδομένα της εφαρμογής.

Κατάταξη	Τίτλος	Impact Factor
1	JAMA : the journal of the American Medical Association	31.718
2	Nature	31.434
3	Cell	31.253
4	Nature reviews. Cancer	30.762
5	Nature genetics	30.259
6	Science (New York, N.Y.)	28.103
7	Nature medicine	27.553
8	Nature immunology	25.113
9	Cancer cell	24.962
10	Nature biotechnology	22.297

Πίνακας 4.2: Τα 10 πρώτα από το σύνολο των περιοδικών στα οποία είναι δημοσιευμένες οι εργασίες της εφαρμογής με τις υψηλότερες τιμές του Impact Factor, όπως αντλήθηκαν από το ετήσιο περιοδικό *Journal Citation Reports* του ιδρύματος Thomson Reuters [10] για το έτος 2008. Σε αυτήν την περίπτωση οι τιμές του Impact Factor είναι υπολογισμένες με δεδομένα εισόδου τις δημοσιεύσεις 8.411 επιστημονικών περιοδικών και τις μεταξύ τους παραπομπές.

Παρατηρούμε λοιπόν ότι οι τιμές της μετρικής που υπολογίστηκαν είναι κατά πολύ μικρότερες από αυτές που παρουσιάζονται στο Διαδίκτυο. Ο τύπος του Impact Factor είναι:

$$\mathbf{IF}(v, y : t) = \frac{Cited(\cup_{i=1..t} V_{y-i}, y)}{|\cup_{i=1..t} V_{y-i}|} \tag{4.7}$$

V_y το σύνολο των άρθρων που δημοσιεύτηκαν στο περιοδικό v το έτος y,
t το χρονικό διάστημα μέσα στο οποίο ανήκει η ένωση των συνόλων V_y
(συνήθως ισούται με 2),
Η συνάρτηση $Cited(A, y)$ μετράει τις αναφορές προς τα άρθρα του συνόλου A των άρθρων που δημοσιεύτηκαν το έτος y.

Με βάση τον τύπο 4.7, οι μικρότερες τιμές της μετρικής συνεπάγονται μικρότερο πλήθος παραπομπών μεταξύ των δημοσιεύσεων των περιοδικών σχετικά με το πλήθος των δημοσιεύσεων. Με άλλα λόγια, ο υπογράφος που δημιουργείται με κόμβους τα περιοδικά και ακμές τις παραπομπές των δημοσιεύσεων της εφαρμογής, είναι αραιός. Επιπλέον, εφόσον οι τιμές του Impact Factor όταν υπολογίζονται στον πλήρη γράφο είναι αρκετά μεγαλύτερες, υπάρχει μεγάλο πλήθος εισερχόμενων ακμών από τον πλήρη γράφο στο υποσύνολο που χρησιμοποιούμε.

Με βάση τα παραπάνω και δεδομένου ότι στους υλοποιημένους αλγόριθμους οι τιμές του Impact Factor των περιοδικών χρησιμοποιήθηκαν κανονικοποιημένες, η αξιολόγηση πρέπει να γίνει με βάση τη θέση που καταλαμβάνει στην κατάταξη κάθε περιοδικό και όχι με βάση την τιμή της μετρικής. Επομένως για τη σύγκριση της υλοποιημένης κατάταξης των περιοδικών, με αυτήν που παρουσιάζεται στο Διαδίκτυο χρησιμοποιήθηκε ο συντελεστής συσχέτισης Spearman:

top-n	Spearman's rho	top-n	Spearman's rho	top-n	Spearman's rho
10	0.648790	110	0.776942	210	0.7042312
20	0.809117	120	0.781027	220	0.6767370
30	0.814130	130	0.779537	230	0.6613285
40	0.802300	140	0.776416	240	0.6500074
50	0.792291	150	0.764033	250	0.6232561
60	0.781503	160	0.755806	260	0.6005150
70	0.764019	170	0.745668	270	0.5645754
80	0.749534	180	0.736873	280	0.5325686
90	0.771130	190	0.7295575	290	0.4941305
100	0.764175	200	0.7165185		

Πίνακας 4.3: Αποτελέσματα του συντελεστή συσχέτισης της κατάταξης των περιοδικών με βάση τις τιμές του Impact Factor 2008, με την κατάταξή τους με βάση της τιμές τις ίδιας μετρικής όπως αντλήθηκαν από το Διαδίκτυο [10]

.

Οι δύο κατατάξεις εμφανίζουν ικανοποιητικό βαθμό ομοιότητας για τα πρώτα 20 έως 60 περιοδικά. Ο συντελεστής συσχέτισης μειώνεται όσο στη σύγκριση προστίθενται περιοδικά που

βρίσκονται χαμηλότερα στην κατάταξη. Αυτό συμβαίνει διότι για τον υπολογισμό του Impact Factor κάθε περιοδικού δεν έχει χρησιμοποιηθεί το σύνολο των δημοσιεύσεών του αλλά αυτές που είναι καταχωρημένες στην εφαρμογή. Συνεπώς, η τελική θέση στην κατάταξη ενός μη δημοφιλούς περιοδικού μπορεί να διαφοροποιηθεί αρκετά από την εισαγωγή των παραπομπών προς μία επιπλέον δημοσίευσή του στον υπολογισμό του Impact Factor σε σχέση με την αντίστοιχη εισαγωγή των παραπομπών μιας δημοσίευσης ενός δημοφιλούς περιοδικού. Αξίζει επίσης να σημειωθεί, ότι τα περιοδικά που βρίσκονται μετά τη $211^{η}$ θέση της κατάταξης, όπως υπολογίστηκε στα πλαίσια της εργασίας, έχουν μηδενική τιμή Impact Factor.

Τέλος, παρ' όλο που η κατάταξη των περιοδικών που δημιουργήθηκε με βάση τον υπολογισμό του Impact Factor διαφέρει από την κατάταξη που έχουν τα περιοδικά στο Διαδίκτυο, δεν μπορούμε να καταλήξουμε στο ότι είναι ανακριβής. Ένα περιοδικό αρκετά δημοφιλές στον ευρύτερο κλάδο της βιολογίας, μπορεί να μην είναι εξίσου σημαντικό στον κλάδο των βιομορίων. Με άλλα λόγια η υλοποιημένη κατάταξη είναι πιο εξειδικευμένη στο επιστημονικό πεδίο των δημοσιεύσεων της εφαρμογής Diana Mirpub συγκριτικά με αυτή του Διαδικτύου.

4.3 Περιγραφή μεθόδου αξιολόγησης

Για την αξιολόγηση των τιμών του Impact Factor, χρησιμοποιήθηκαν οι τιμές της λίστας που αναρτάται στο ετήσιο περιοδικό Journal Citation Reports σαν «πραγματικές τιμές» της μετρικής. Αντίστοιχη επίσημη λίστα με τα σκορ του αλγόριθμου PageRank του συνόλου των επιστημονικών εργασιών των οποίων το περιεχόμενο είναι διαθέσιμο στο Διαδίκτυο, δε βρέθηκε. Όποια εργαλεία έχουν δημιουργηθεί που υπολογίζουν το PageRank αφορούν κυρίως σε ιστοσελίδες και ο υπολογισμός βασίζεται στους συνδέσμους μεταξύ των ιστοσελίδων αυτών. Κρίνεται λοιπόν απαραίτητη η δημιουργία μιας μεθόδου αξιολόγησης όπου η «πραγματική κατάταξη» που θα αποτελεί το μέτρο σύγκρισης των αποτελεσμάτων των υλοποιημένων αλγορίθμων, θα προκύπτει από τα δεδομένα που ήδη έχουμε στη διάθεσή μας.

Όπως έχει αναφερθεί σε προηγούμενα κεφάλαια, ο κύριος λόγος που καθιστά τα σκορ του αλγόριθμου PageRank μη αξιόπιστα για την αξιολόγηση των δημοσιεύσεων, είναι ότι τη δεδομένη χρονική στιγμή που υπολογίζονται τα σκορ, δεν είναι γνωστό το πλήθος των αναφορών που πρόκειται να γίνουν προς τις δημοσιεύσεις στο μέλλον. Αυτή η κατάσταση επηρεάζει κυρίως την αξιολόγηση των νέων επιστημονικών εργασιών εφόσον δεν έχει περάσει αρκετός καιρός από τη δημοσίευσή τους, διότι ακόμα και αν ειναι αξιόλογες, πρέπει να περάσει κάποιος χρόνος εως ότου γίνουν αναφορές προς αυτές. Αντίθετα, οι παλαιότερες εργασίες
έχουν συγκεντρώσει επαρκές σύνολο παραπομπών, το οποίο ακόμα και αν συνεχίσει να αυξάνεται θα είναι με μειούμενο ρυθμό, άρα θεωρούμε ότι το PageRank τις κατατάσσει σχετικά σωστά.

Επομένως, σαν «πραγματικές τιμές» των σκορ κατάταξης των δημοσιεύσεων θα μπορούσαμε να ορίσουμε τα σκορ του αλγόριθμου PageRank που πρόκειται να λάβουν οι δημοσιεύσεις στο μέλλον. Όπως παρουσιάζεται στο Σχήμα 2.9 του $2^{ου}$ Κεφαλαίου, το μεγαλύτερο πλήθος παραπομπών ανά έτος προς μία επιστημονική εργασία δημιουργείται 2 χρόνια μετά τη δημοσίευσή της, ενώ στη συνέχεια μειώνεται.

Με βάση τα παραπάνω, τα βήματα της μεθόδου αξιολόγησης που χρησιμοποιήθηκε είναι τα εξής:

1. Εφαρμογή του αλγόριθμου PageRank στο υποσύνολο των επιστημονικών εργασιών της εφαρμογής Diana Mirpub που είναι δημοσιευμένες από το έτος 1994 (χρονολογία την οποία είναι δημοσιευμένες οι πιο παλιές εργασίες της εφαρμογής) μέχρι και το έτος 2011.

2. Δημιουργία της λίστας με τις «πραγματικές τιμές» του αλγόριθμου PageRank κρατώντας από τις τιμές του βήματος 1 μόνο αυτές που αφορούν εργασίες δημοσιευμένες μέχρι και το έτος 2008.

3. Εφαρμογή του αλγόριθμου PageRank και των υλοποιημένων αλγόριθμων στο υποσύνολο των επιστημονικών εργασιών της εφαρμογής Diana Mirpub που είναι δημοσιευμένες στο διάστημα 1994 έως και 2008.

4. Υπολογισμός του συντελεστή συσχέτισης Spearman μεταξύ των κατατάξεων που προκύπτουν από τα αποτελέσματα του βήματος 3 και αυτής που ορίσαμε ως «πραγματική κατάταξη» (Βήμα 2).

4.4 Αποτελέσματα - Συμπεράσματα

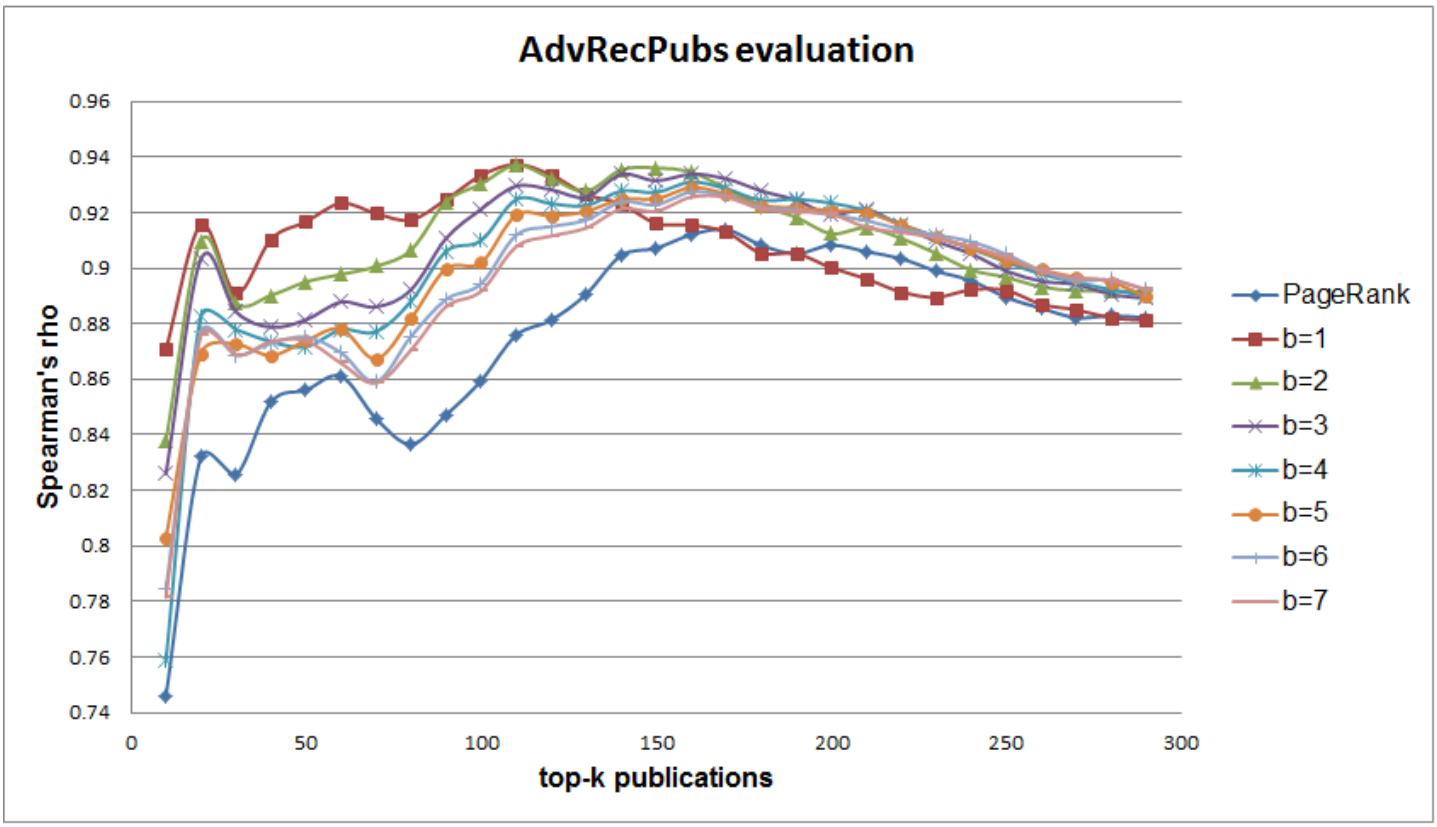
Παρακάτω αξιολογούνται τα αποτελέσματα των υλοποιημένων αλγόριθμων με την εφαρμογή της μεθόδου που περιγράψαμε. Εξετάζεται κατά πόσο οι κατατάξεις των αλγορίθμων προσεγγίζουν το PageRank σε μελλοντική χρονική στιγμή, στην οποία θεωρούμε ότι δίνει την πραγματική κατάταξη. Γι' αυτό το λόγο, υπολογίστηκε ο συντελεστής συσχέτισης μεταξύ των αποτελεσμάτων κάθε αλγόριθμου με την «πραγματική κατάταξη».

Όταν ένας χρήστης πραγματοποιεί ένα ερώτημα σε μια μηχανή αναζήτησης, ενδιαφέρεται κυρίως για τα πρώτα αποτελέσματα που επιστρέφονται από αυτή. Είναι επιθυμητό λοιπόν, οι κατατάξεις - αποτελέσματα των υλοποιημένων αλγόριθμων να εμφανίζουν παρόμοια αποτελέσματα με αυτά της «πραγματικής», κυρίως για τις δημοσιεύσεις που καταλαμβάνουν τις πρώτες θέσεις σε αυτές. Επομένως, πρέπει να αξιολογηθούν και τα αποτελέσματα της κάθε κατάταξης που αφορούν στις 10,20 κ.ο.κ. πρώτες δημοσιεύσεις σύμφωνα με αυτήν, και όχι μόνο το σύνολό της. Γι' αυτό το λόγο ο συντελεστής συσχέτισης υπολογίστηκε για διάφορα υποσύνολα των κατατάξεων, το πρώτο στοιχείο των οποίων ήταν πάντα η δημοσίευση που βρίσκεται στην πρώτη θέση και το τελευταίο αυτή που βρίσκεται στη θέση k. Όσο η τιμή του συντελεστή συσχέτισης προσεγγίζει τη μονάδα τόσο πιο κοντά στην κατατάξη που έχουμε ορίσει ως «πραγματική» βρίσκονται τα αποτελέσματα.

4.4.1 Αξιολόγηση αλγόριθμου AdvRecPubs

Στο διάγραμμα του Σχήματος 4.5 συγκρίνεται η κατάταξη του PageRank στο πλήρες σύνολο, με αυτή του AdvRecPubs λαμβάνοντας υπόψη το γράφο που έχουμε με τις δημοσιεύσεις μέχρι και το 2008. Η σύγκριση πραγματοποιείται για διάφορες τιμές του k σε top-k αποτελέσματα. Παρατηρούμε ότι όσο μικρότερη είναι η τιμή της παραμέτρου b, τόσο περισσότερο

η κατάταξη των 110 πρώτων δημοσιεύσεων προσεγγίζει την «πραγματική» (ο συντελεστής προσεγγίζει τη μονάδα). Το αντίθετο συμβαίνει όταν στον υπολογισμό του συντελεστή συσχέτισης προστίθενται δημοσιεύσεις που βρίσκονται μετά την $110^{η}$ θέση της κατάταξης. Με βάση τον τύπο 3.1 όσο μειώνεται η τιμή της παραμέτρου b, τόσο περισσότερο προωθούνται οι νεότερες δημοσιεύσεις σε βάρος των παλαιότερων. Συμπερασματικά, όταν προωθούμε τις νέες δημοσιεύσεις, αυτές που έχουν ήδη καλές θέσεις στην κατάταξη, καταλαμβάνουν θέσεις ακόμα πιο κοντά στις «πραγματικές», ενώ η τελική θέση των δημοσιεύσεων που βρίσκονται αρκετά χαμηλά, δεν βελτιώνεται συγκριτικά με αυτή στην οποία τις κατατάσσει ο αλγόριθμος PageRank.



Σχήμα 4.5: Ο συντελεστής συσχέτισης Spearman του αλγόριθμου AdvRecPubs με την «πραγματική κατάταξη», για διάφορες τιμές της παραμέτρου b. Με βάση τα αποτελέσματα, η κατάταξη των δημοσιεύσεων προσεγγίζει περισσότερο την «πραγματική» συνολικά, όταν $b = 3$.

Παρακάτω παρουσιάζεται το πλήθος των εργασιών που εμφανίζονται σε κάθε υποσύνολο της κατάταξης, ομαδοποιημένες σύμφωνα με τη χρονολογία δημοσίευσής τους.

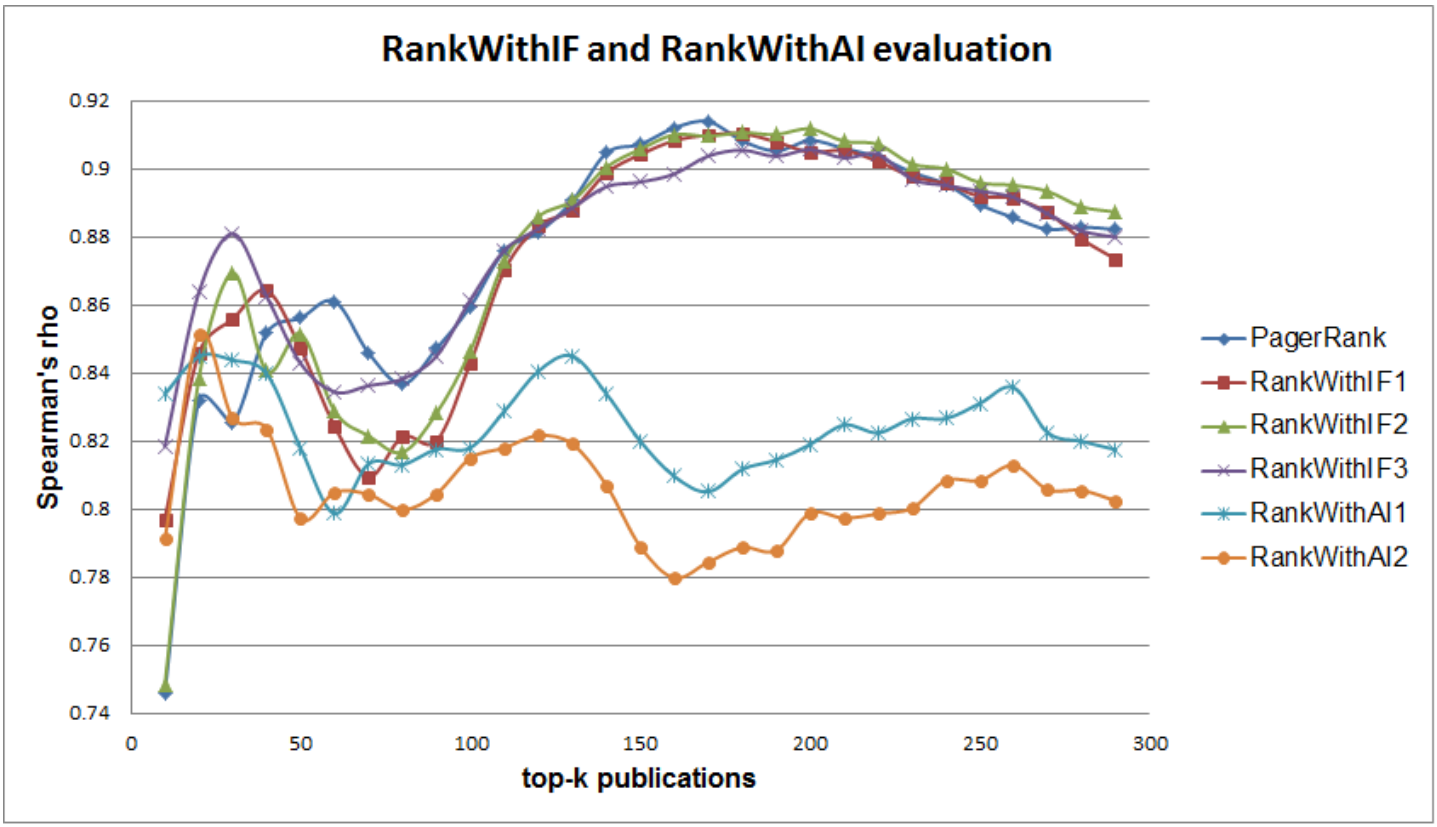
	$1994-2000$	$2001-2006$	$2007-2008$
0 - 50	3	40	7
50 - 150	4	69	27
150 - 1834	17	404	1263
Σύνολο	24	513	1297

Πίνακας 4.4: Το πλήθος των εργασιών ανάλογα με το έτος δημοσίευσης τους και τη θέση που καταλαμβάνουν στην τελική κατάταξη του αλγόριθμου PageRank.

Όπως παρατηρούμε, το 71% των εργασιών στις οποίες εφαρμόζουμε τους υλοποιημένους αλγόριθμους είναι δημοσιευμένες τα έτη 2007-2008, το 97% των οποίων κατατάσσονται σε θέσεις μεγαλύτερες της $150^{ης}$, σύμφωνα με τον αλγόριθμο PageRank. Καταλήγουμε λοιπόν ότι ο υλοποιημένος αλγόριθμος δεν εμφανίζει βελτιωμένα αποτελέσματα συγκριτικά με τον PageRank για τις δημοσιεύσεις που κατατάσσονται χαμηλά στη λίστα, διότι προωθεί το σύνολό τους με τον ίδιο τρόπο.

4.4.2 Αξιολόγηση αλγορίθμων RankWithIF και RankWithAI

Από το διάγραμμα του σχήματος 4.6, συμπεραίνουμε ότι η παραμετροποίηση του αλγόριθμου PageRank ώστε να προωθούνται οι επιστημονικές εργασίες με βάση τη δημοτικότητα (Impact Factor) ή το κύρος (Article Influence) του αντίστοιχου περιοδικού, δε βελτίωσε αισθητά την απόδοσή του. Ιδιαίτερα στην περίπτωση που χρησιμοποιήθηκε ο αλγόριθμος Article Influence για την αξιολόγηση των περιοδικών, τα αποτελέσματα προσεγγίζουν ακόμα λιγότερο την «πραγματική κατάταξη» από αυτά του PageRank.



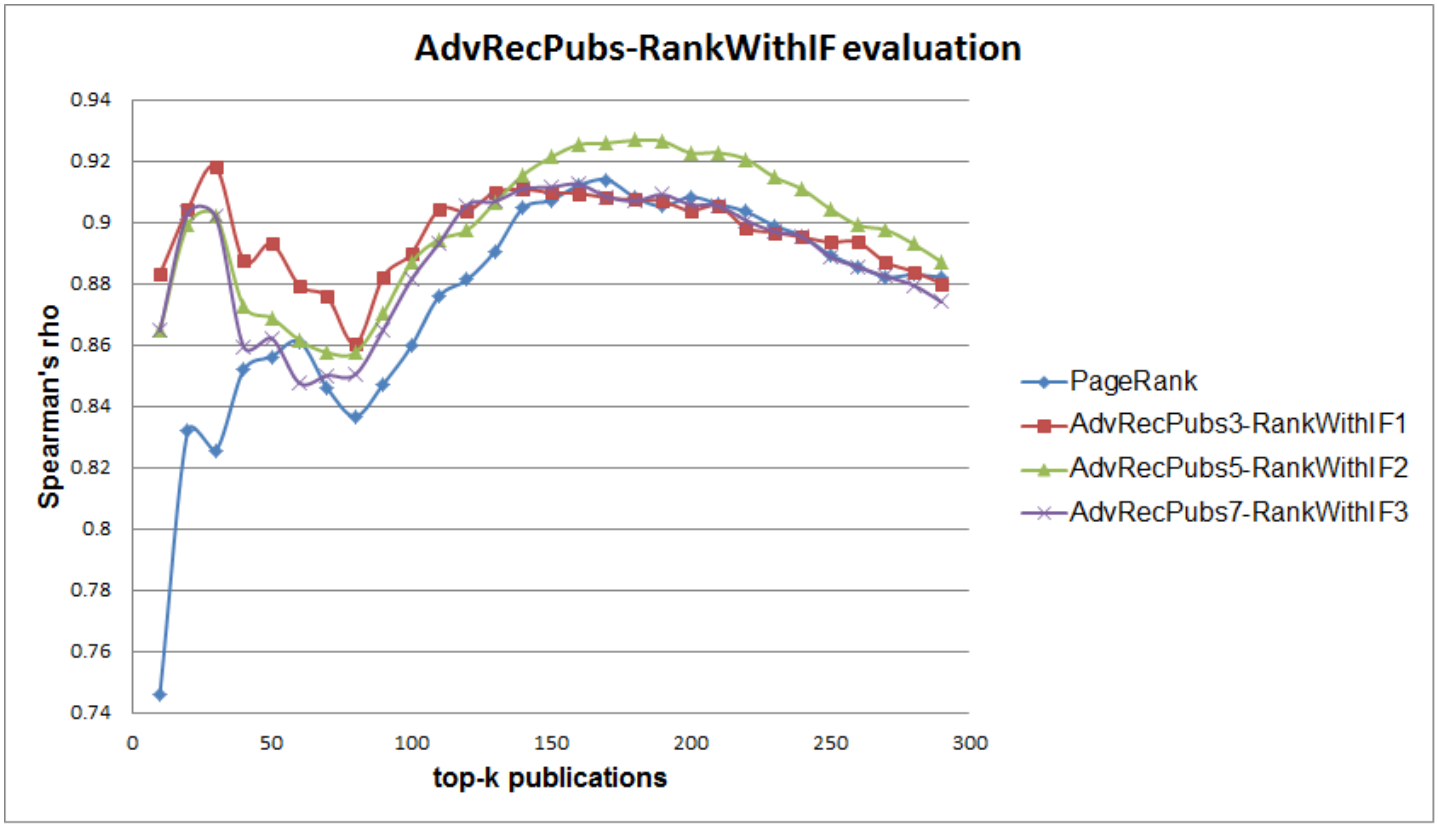
Σχήμα 4.6: Αποτελέσματα αλγόριθμων κατάταξης που χρησιμοποιούν τη δημοτικότητα/κύρος του αντίστοιχου περιοδικού όπως περιγράφηκαν στην Ενότητα 3.1.3.

Όπως περιγράφηκε στην Ενότητα 4.2, για να βγάλουμε συμπεράσματα για τη δημοτικότητα των περιοδικών, χρησιμοποιήσαμε μόνο τις τις επιστημονικές εργασίες της εφαρμογής Diana Mirpub και όχι το σύνολο εργασιών που είναι δημοσιευμένες στα περιοδικά αυτά, χωρίς να γνωρίζουμε αν το σύνολο των εργασιών που διαθέτουμε για κάθε περιοδικό είναι αντιπροσωπευτικό. Ο λόγος που ο αλγόριθμος ο οποίος χρησιμοποιεί το κύρος των περιοδικών με βάση το σκορ Article Influence, έχει ακόμα χειρότερα αποτελέσματα θα μπορούσε να είναι ότι βασίζεται ακόμα περισσότερο στην τοπολογία του γράφου που δημιουργείται, με κόμβους τα εξεταζόμενα περιοδικά και ακμές τις δημοσιεύσεις της εφαρμογής. Ο γράφος αυτός είναι αρκετά αραιός. Παρά όλα αυτά, συγκριτικά με τους υπόλοιπους αλγόριθμους, θεωρούμε ότι αυτός που χρησιμοποιεί τα αποτελέσματα του Impact Factor υπολογισμένο για διαφορετικό έτος για

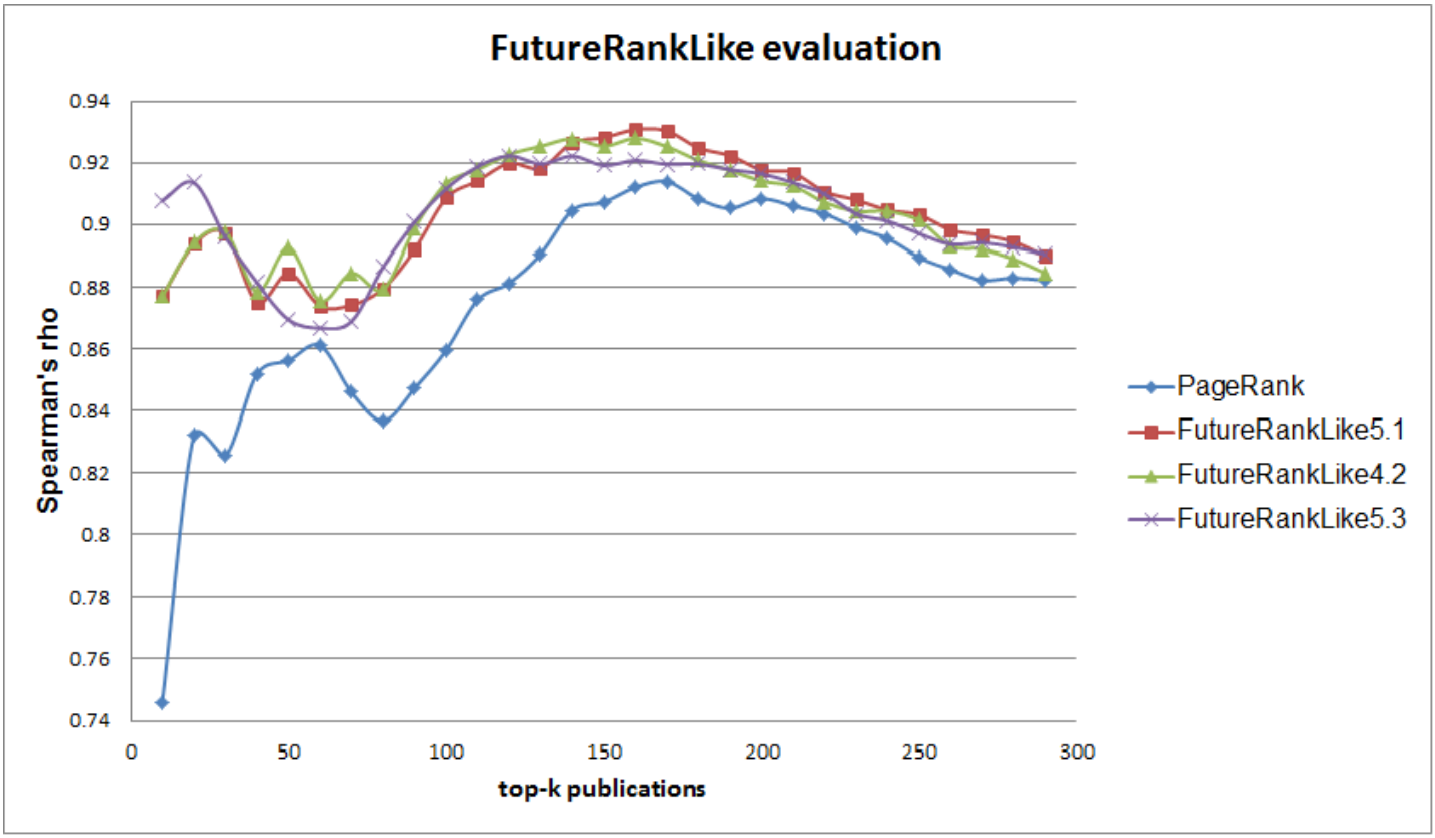
κάθε δημοσίευση, έχει τα καλύτερα αποτελέσματα, κυρίως για τις πρώτες 60 δημοσιεύσεις.

4.4.3 Αξιολόγηση αλγορίθμων AdvRecPubs-RankWithIF και FutureRankLike

Στις ενότητες 3.1.4 και 3.1.6 περιγράψαμε τους αλγόριθμους AdvRecPubs-RankWithIF και FutureRankLike αντίστοιχα. Η διαφορά έγκειται στον τρόπο με τον οποίο συνδυάζουν την προώθηση των νέων εργασιών σε σχέση με την προώθηση αυτών που είναι δημοσιευμένες σε αξιόλογα περιοδικά. Στο σημείο αυτό, παρουσιάζεται η διαφορά στη συμπεριφορά του «τυχαίου περιηγητή»: Στην πρώτη περίπτωση ο τυχαίος περιηγητής προτιμά εργασίες που ειναι ταυτόχρονα νεες, αλλά και δημοσιευμένες σε δημοφιλή περιοδικά. Αντίθετα, στη δεύτερη περίπτωση, προτιμά εργασίες που είναι είτε δημοσιευμένες σε δημοφιλή περιοδικά, είτε είναι νέες, είτε και τα δύο.



Σχήμα 4.7: Αποτελέσματα αλγόριθμου AdvRecPubs-RankWithIF για διαφορετικούς τρόπους υπολογισμού του impact factor και τιμές της σταθεράς b.



Σχήμα 4.8: Αποτελέσματα του αλγόριθμου FutureRankLike για διαφορετικούς τρόπους υπολογισμού του impact factor και τιμές της σταθεράς b.

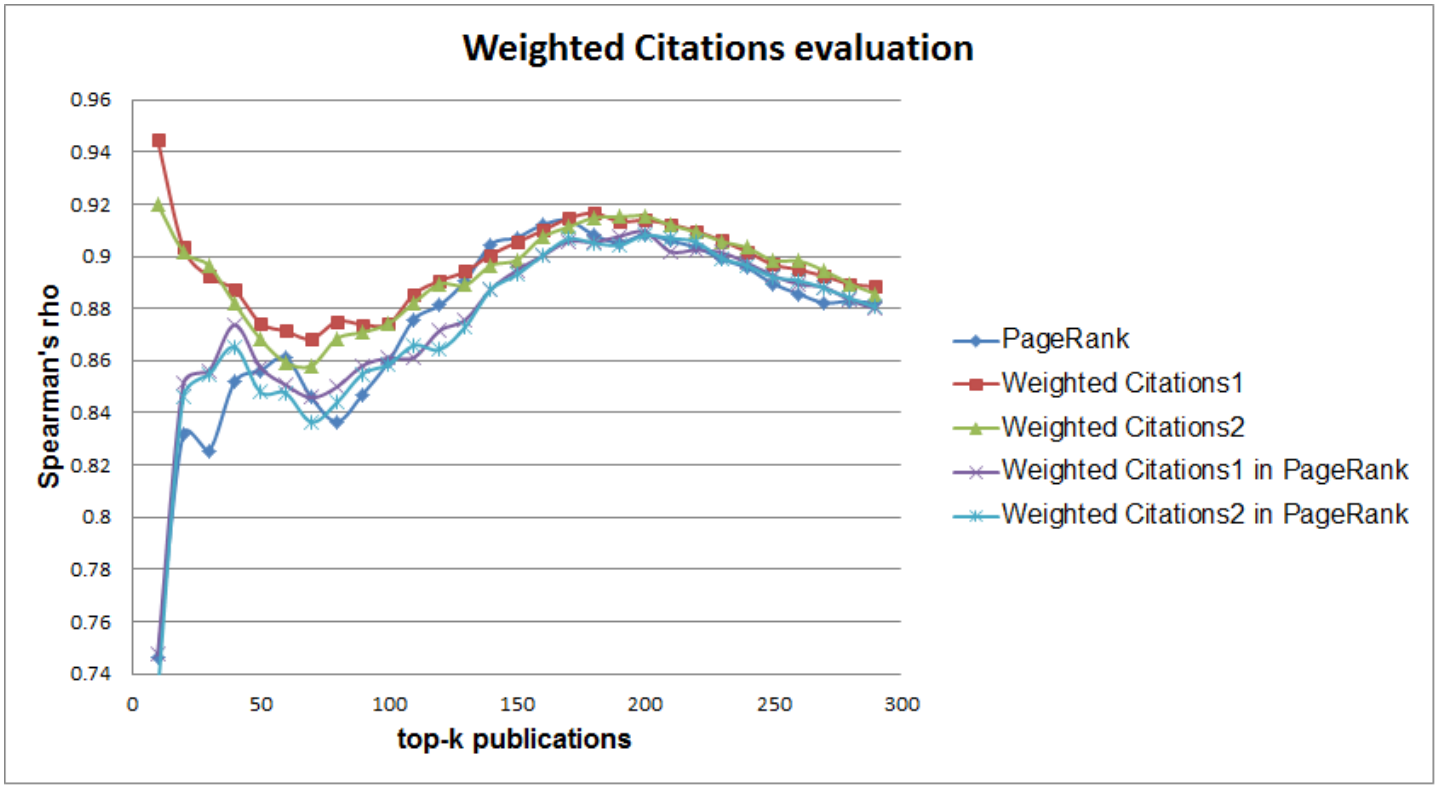
Είναι επομένως αναμενόμενο οι αλγόριθμοι να εμφανίζουν καλύτερα αποτελέσματα σε σχέση με αυτούς των δύο προηγούμενων ενοτήτων αφού συνδυάζουν τους μηχανισμούς προώθησης των δημοσιεύσεων που αυτοί χρησιμοποιούν. Ενδεικτικά, το πρόβλημα προώθησης με τον ίδιο τρόπο όλων των εργασιών που καταλαμβάνουν χαμηλές θέσεις με βάση την κατάταξη του αλγόριθμου PageRank, όπως περιγράφηκε στην Ενότητα 4.4.1, αντιμετωπίζεται όταν στα τελικά σκορ συμπεριλαμβάνεται και η προώθηση των εργασιών με βάση τη δημοτικότητα των περιοδικών.

4.4.4 Αξιολόγηση αλγόριθμου Weighted Citations

Ο αλγόριθμος που περιγράφεται στην Ενότητα 3.1.5 είναι ο μοναδικός που δεν αποτελεί μία παραμετροποίηση του PageRank. Είναι αναμενόμενο τα αποτελέσματά του να έχουν τις λιγότερες ομοιότητες με αυτά του PageRank συγκριτικά με τους υπόλοιπους αλγόριθμους.

Σύμφωνα με το Σχήμα 4.9, ιδιαίτερα στις 50 πρώτες δημοσιεύσεις, ο συντελεστής συσχέτισης της κατάταξης του συγκεκριμένου αλγόριθμου με την «πραγματική», έχει εντελώς διαφορετική συμπεριφορά από τον αντίστοιχο συντελεστή του αλγόριθμου PageRank, η οποία είναι εμφανώς καλύτερη. Έγινε προσπάθεια αυτή η συμπεριφορά να βελτιωθεί και για τις δημοσιεύσεις που βρίσκονται χαμηλότερα στην κατάταξη. Δεδομένου ότι ο αλγόριθμος PageRank δίνεται από τον τύπο 2.3, το διάνυσμα με όλα τα στοιχεία του ίσα με τη μονάδα,

αντικαταστάθηκε με τα αποτελέσματα του συγκεκριμένου αλγόριθμου. Όπως παρατηρούμε η κατάταξη που προκύπτει είναι απόλυτα συσχετισμένη με την αντίστοιχη του αλγόριθμου PageRank σε σημείο που να μην έχει σημαντική διαφορά από αυτή.

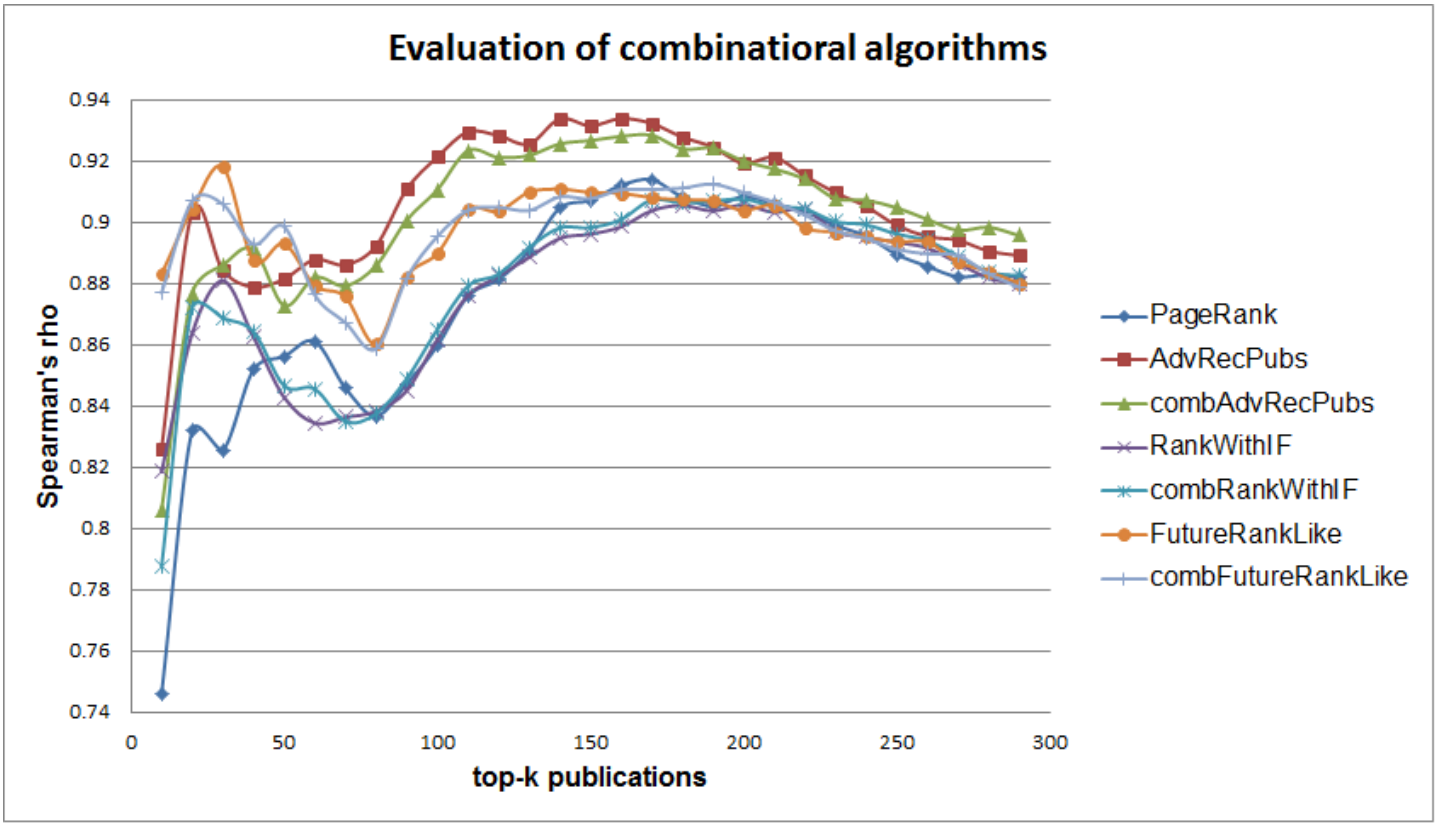


Σχήμα 4.9: Αποτελέσματα αλγόριθμου Weighted Citations με διαφορετικούς τρόπους υπολογισμού της μετρικής EigenFactor μαζί με τα αποτελέσματα του συνδυαστικού αλγόριθμου Weighted Citations in Pagerank.

4.4.5 Αξιολόγηση συνδυαστικού μηχανισμού κατάταξης

Όπως έχει ήδη αναφερθεί, το σύνολο των δημοσιεύσεων των οποίων η δημοτικότητα δεν ανταποκρίνεται στην αξία τους, είναι αυτές που έχουν δημοσιευθεί πρόσφατα. Γι' αυτό το λόγο, υλοποιήσαμε συνδυαστικό αλγόριθμο που χρησιμοποιεί τόσο τα αποτελέσματα του PageRank όσο και αυτά ενός από τους αλγόριθμους που εξετάσαμε. Το τελικό σκορ κάθε εργασίας υπολογίζεται από τον συνδυασμό των σκορ που δίνουν οι παραπάνω αλγόριθμοι ανάλογα με το έτος δημοσίευσής της. Συγκεκριμένα, για τις εργασίες που είναι δημοσιευμένες τα έτη 2007 και 2008, τα σκορ κατάταξης δίνονται εξ' ολοκλήρου από τον υλοποιημένο αλγόριθμο που εξετάζουμε κάθε φορά, για τις εργασίες που είναι δημοσιευμένες τα έτη 2005 και 2006, τα σκορ αποτελούνται από το άθροισμα του 5% του PageRank και του 95% του υλοποιημένου αλγόριθμου, για τις εργασίες που είναι δημοσιευμένες τα έτη 2003 και 2004, τα σκορ είναι το 10% του PageRank συν το 90% του υλοποιημένου αλγόριθμου, κ.ο.κ. μέχρι το έτος 1993, το οποίο αποτελεί το έτος δημοσίευσης των πιο παλιών εργασιών, όπου τα σκορ των εργασιών είναι το άθροισμα του 40% του PageRank συν του 60% του υλοποιημένου αλγόριθμου.

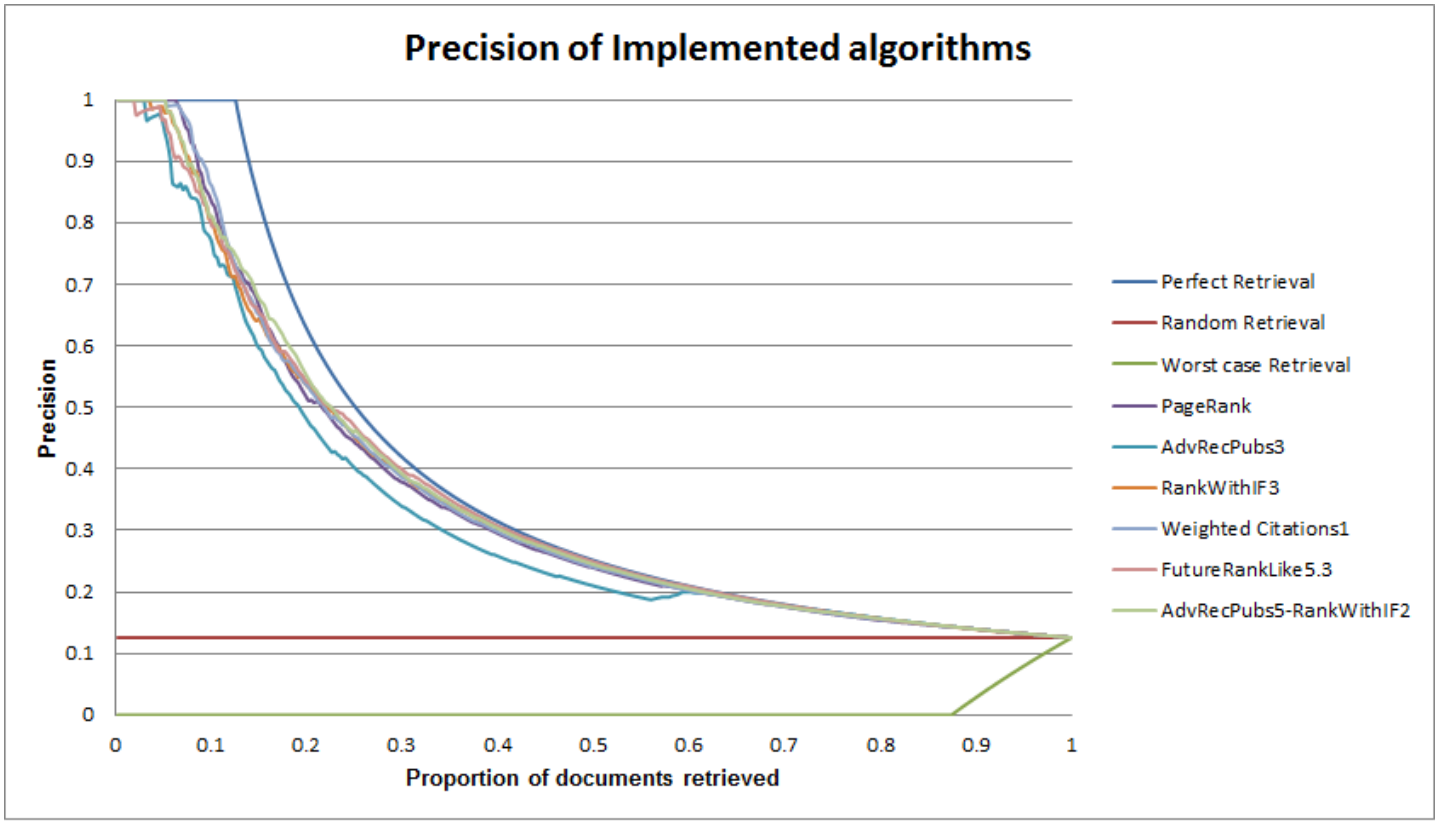
Όμως, η πλειοψηφία των εργασιών που διαθέτουμε είναι δημοσιευμένες τα έτη 2007-2008, με αποτέλεσμα ο συνδυαστικός αλγόριθμος να μην έχει ιδιαίτερη αξία τουλάχιστον για το συγκεκριμένο σύνολο δημοσιεύσεων.



Σχήμα 4.10: Αποτελέσματα της ποσοστιαίας χρήσης των αποτελεσμάτων των υλοποιημένων αλγορίθμων όπως περιγράφονται στην Ενότητα 3.1.7.2.

4.4.6 Γενικά συμπεράσματα - επιλογή βέλτιστου μηχανισμού

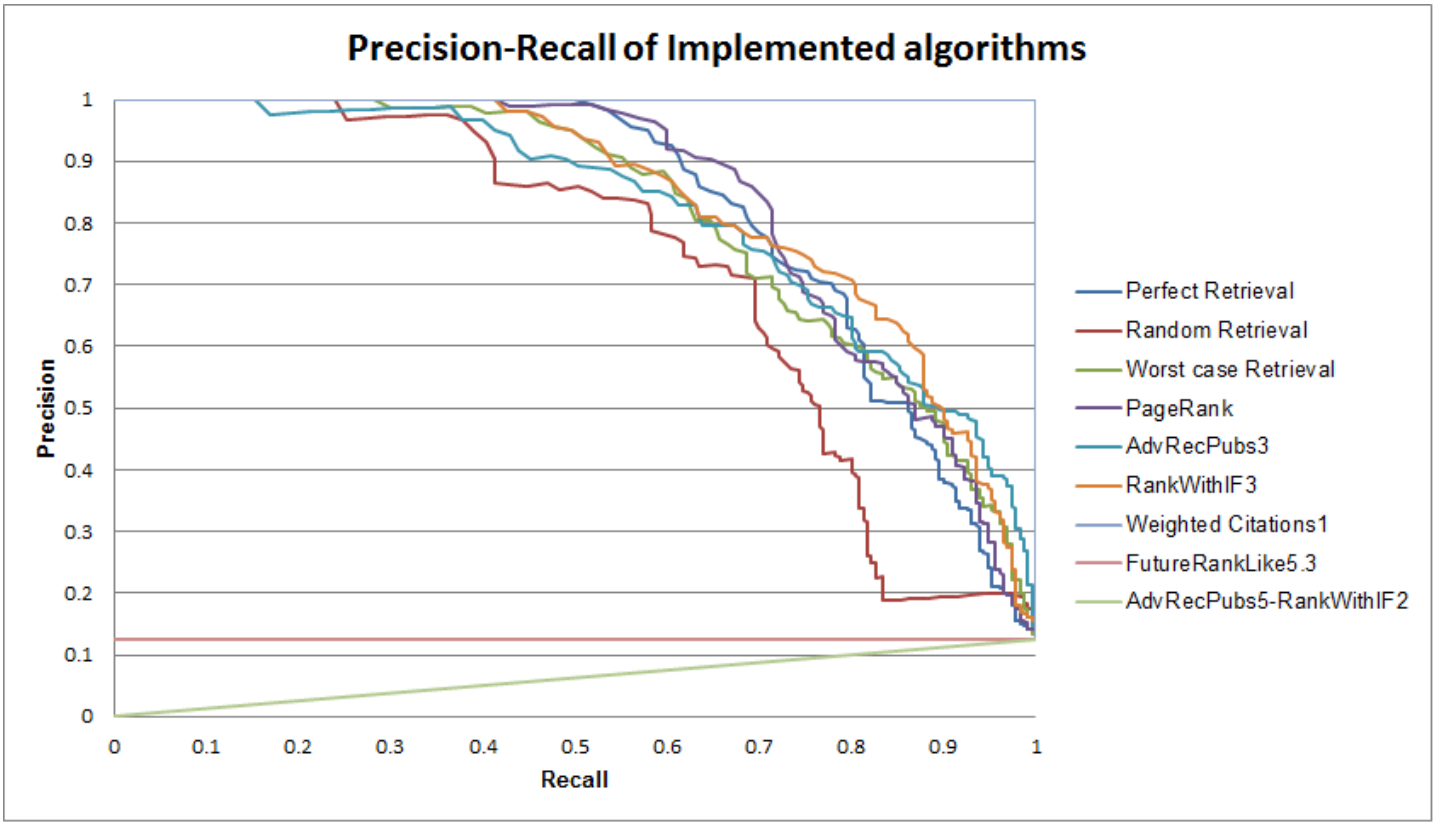
Όπως έχουμε επισημάνει, όταν πραγματοποιείται κάποιο ερώτημα σε μία μηχανή αναζήτησης, τα πρώτα από τα αποτελέσματα πρέπει να είναι αυτά που ενδιαφέρουν τον χρήστη. Γι' αυτό το λόγο ο συντελεστής συσχέτισης υπολογίστηκε μόνο με βάση τις top-k δημοσιεύσεις των κατατάξεων-αποτελεσμάτων των αλγορίθμων. Δεδομένου ότι η συμπεριφορά των υλοποιημένων αλγόριθμων ποικίλει στις διάφορες θέσεις των δημοσιεύσεων στην κατάταξη, για να μπορέσουμε να τους συγκρίνουμε, πρέπει να ορίσουμε τη θέση της κατάταξης μέχρι την οποία μας ενδιαφέρει η αποτελεσματικότητά τους. Με βάση τα αποτελέσματα των αλγόριθμων, το μεγαλύτερο μέρος του αθροίσματος των σκορ PageRank (που συνολικά αθροίζουν στο 1), για το δεδομένο σύνολο, βρίσκεται στις 230 δημοσιεύσεις της κάθε κατάταξης και γι' αυτό επιλέχθηκε σαν σημείο αναφοράς. Γι' αυτό το λόγο ο συντελεστής συσχέτισης υπολογίστηκε με βάση τις top-k δημοσιεύσεις με μέγιστο $k = 230$.



Σχήμα 4.11: Η μετρική της ακρίβειας έναντι του ποσοστού των ανακτημένων δημοσιεύσεων. Το ποσοστό 100% αντιστοιχίζεται με το σύνολο των δημοσιεύσεων που είναι δημοσιευμένες ανάμεσα στα έτη 1994-2008 (1834).

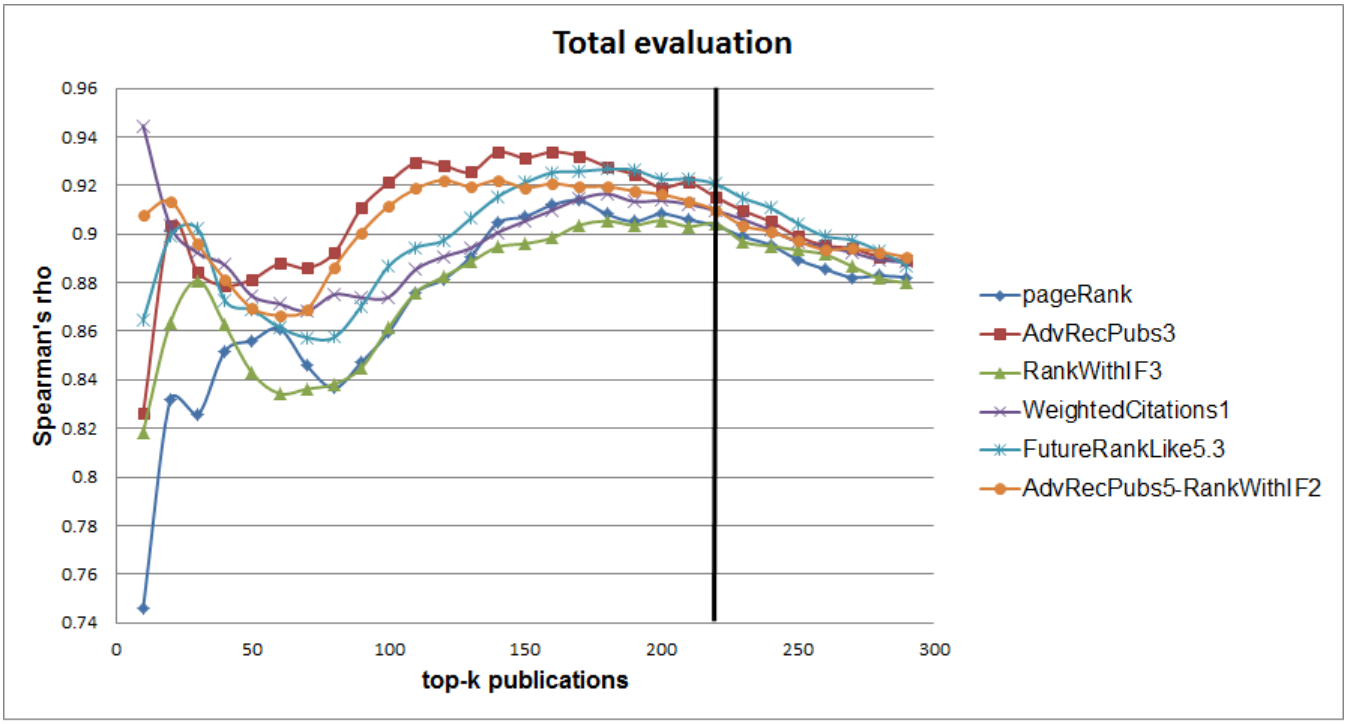
Αν και οι μετρικές της ακρίβειας και της ανάκλησης συνήθως χρησιμοποιούνται για να αξιολογηθεί ο μηχανισμός ανάκτησης της πληροφορίας με βάση τα ερωτήματα που εισάγει ο χρήστης, κατά πόσο δηλαδή τα αποτελέσματα που επιστρέφονται είναι σχετικά με το ερώτημα, εμείς τις χρησιμοποιήσαμε για να αξιολογήσουμε αν οι μηχανισμοί κατάταξης των δημοσιεύσεων που υλοποιήσαμε, έχουν στις 230 πρώτες θέσεις, τις δημοσιεύσεις των αντίστοιχων θέσεων της ιδανικής κατάταξης. Συγκεκριμένα, θεωρούμε ότι πραγματοποιούμε ένα ερώτημα για το οποίο οι σχετικές εργασίες είναι οι πρώτες 230 δημοσιεύσεις της ιδανικής κατάταξης. Εφαρμόζοντας τους τύπους 4.5 και 4.6 της ανάκλησης και της ακρίβειας αντίστοιχα, παίρνουμε τα αποτελέσματα του Διαγράμματος 4.11.

Στόχος οποιουδήποτε αλγόριθμου κατάταξης είναι να βελτιώνει τα αποτελέσματα της «τυχαίας» και να προσεγγίζει αυτά της «ιδανικής» κατάταξης, κάτι που οι υλοποιημένοι αλγόριθμοι φαίνεται να πετυχαίνουν.



Σχήμα 4.12: Οι μετρικές της ακρίβειας έναντι της ανάκλησης για τους υλοποιημένους αλγόριθμους.

Στο Διάγραμμα 4.13 εμφανίζονται τα συγκεντρωτικά αποτελέσματα των αλγόριθμων κατάταξης που περιγράψαμε. Για να καταλήξουμε στους βέλτιστους δυνατούς αλγόριθμους, συγκρίνουμε τις καμπύλες του συντελεστή συσχέτισης των αντίστοιχων κατατάξεων για τις 230 πρώτες δημοσιεύσεις.



Σχήμα 4.13: Σύγκριση των υλοποιημένων αλγόριθμων αποτιμώντας τις παραμέτρους με βάση τα αποτελέσματα των προηγούμενων συγκρίσεων.

Παρατηρούμε ότι όσο προστίθενται δημοσιεύσεις στον υπολογισμό του συντελεστή συσχέτισης, αυτός συνεχώς βελτιώνεται μέχρι οι δημοσιεύσεις να φτάσουν τις πρώτες 150, ενώ στη συνέχεια παραμένει σταθερός. Όπως είδαμε δε σε προηγούμενα διαγράμματα της ενότητας, ο συντελεστής συσχέτισης μειώνεται όταν στον υπολογισμό του προστίθενται δημοσιεύσεις με θέσεις μεγαλύτερες της $230^{ης}$. Αυτό το αποδίδουμε στο γεγονός ότι όσο πιο χαμηλά στην κατάταξη βρίσκονται οι δημοσιεύσεις, οι διαφορές των σκορ κατάταξής τους είναι πολύ μικρές, με αποτέλεσμα οι θέσεις τους να είναι πολύ ευαίσθητες σε οποιαδήποτε μεταβολή, φαινόμενο το οποίο δεν μας ενδιαφέρει μιας και εστιάζουμε στις σημαντικότερες δημοσιεύσεις κάθε κατάταξης.

Οι αλγόριθμοι με την καλύτερη συμπεριφορά εμφανίζονται να είναι οι AdvRecPubs3, FutureRankLike5.3 και AdvRecPubs5-RankWithIF2, με τον πρώτο να μην είναι αποτελεσματικός στις 10 πρώτες δημοσιεύσεις της κατάταξης οι οποίες είναι και οι πιο σημαντικές, καθώς και να εμφανίζει τα χειρότερα αποτελέσματα στα διαγράμματα της ακρίβειας και της ανάκλησης. Συμπερασματικά, οι αλγόριθμοι που επιλέγονται για να ενσωματωθούν στην εφαρμογή Diana Mirpub είναι οι FutureRankLike3, AdvRecPubs-RankWithIF2 καθώς και ο αλγόριθμος PageRank σαν βασικός μηχανισμός κατάταξης των δημοσιεύσεων.

Κεφάλαιο 5

Ενσωμάτωση υλοποιημένων αλγορίθμων στην εφαρμογή MirPub

Στο κεφάλαιο αυτό περιγράφεται το μέρος της δομής της εφαρμογής DIANA MirPub που μας ενδιαφέρει, έτσι ώστε να γίνει κατανοητός ο τρόπος ενσωμάτωσης του μηχανισμού κατάταξης των δημοσιεύσεων, ενώ αναλύονται και οι λειτουργίες που πρέπει να εκτελεί ο μηχανισμός αυτός. Στη συνέχεια, παρουσιάζονται τα εργαλεία στα οποία έχει βασιστεί η ανάπτυξη της υπάρχουσας εφαρμογής, τα οποία χρησιμοποιήθηκαν και για την επέκτασή της. Τέλος περιγράφονται οι λεπτομέρειες υλοποίησης του μηχανισμού κατάταξης. Να σημειωθεί ότι η λειτουργικότητα που περιγράφεται σε αυτό το σημείο, αποτέλεσε μέρος της επιστημονικής εργασίας που έγινε δεκτή και παρουσιάστηκε στο συνέδριο TPDL, τον Σεπτέμβριο του 2015 [9]. Η εργασία αυτή, είχε σαν σκοπό την επέκταση της εφαρμογής Diana Mirpub, ώστε να καθίσταται όσο το δυνατόν πιο εύκολη η εύρεση των επιθυμητών εργασιών από τους χρήστες της.

5.1 Δομή της εφαρμογής

Η ιστοσελίδα DIANA είναι υλοποιημένη με τη χρήση του Yii framework της PHP και επικοινωνεί με μια βάση δεδομένων mySQL με όνομα diana_universe. Στη συνέχεια, περιγράφονται οι κλάσεις της εφαρμογής DIANA MirPub και ο πίνακας της βάσης που τροποποιήθηκαν για την υλοποίηση της ζητούμενης λειτουργικότητας.

Ο πίνακας diana_paper

Ο πίνακας αυτός διαθέτει πληροφορίες για τα άρθρα που είναι αποθηκευμένα στην εφαρμογή, όπως τον τίτλο του κάθε άρθρου, τους συγγραφείς του, το έτος και το περιοδικό δημοσίευσης του καθώς και τα ids τους στις βάσεις MEDLINE και PMC.

main.php

Το αρχείο αυτό όπως και σε όλες τις εφαρμογές που είναι βασισμένες στο Yii framework

αποτελεί το configuration αρχείο της εφαρμογής.

MirpubController

Η κλάση αυτή αποτελεί τον ελεγκτή της εφαρμογής DIANA mirPub και διαθέτει μεθόδους (actions) που είναι υπεύθυνες για την επεξεργασία και αναγνώριση των λέξεων-κλειδιών που δίνονται στη φόρμα αναζήτησης, για την προβολή των αποτελεσμάτων που αφορούν microRNAs, την προβολή του γράφου εξέλιξης του ονόματος του αντίστοιχου microRNA, την προβολή διάφορων στατιστικών στοιχείων καθώς και την εισαγωγή νέων δημοσιεύσεων στη βάση. Οι μέθοδοι που μας αφορούν είναι η actionGetResults που χρησιμοποιείται για την παραγωγή της τελικής σελίδας αποτελεσμάτων και η επιμέρους μέθοδος actionResults που καλείται από την προηγούμενη και χρησιμοποιείται για την επεξεργασία των λέξεων-κλειδιών που δίνονται στη φόρμα αναζήτησης, την παραγωγή των επιμέρους αποτελεσμάτων και την προβολή τους.

CitationFetcher Model

Η κλάση CitationFetcher τύπου Model χρησιμοποιείται για τη συλλογή δημοσιεύσεων με βάση κάποιες λέξεις-κλειδιά. Από τις μεθόδους της αυτή που μας ενδιαφέρει είναι η get_citations_by_type_array η οποία χρησιμοποιείται από τον MirpubController.
Είσοδος της μεθόδου αυτής είναι το σύνολο των λέξεων-κλειδιών βάσει των οποίων πρέπει να γίνει η αναζήτηση στη βάση και ο τύπος του microRNA που αντιστοιχεί στην λέξη κλειδί που εισήγαγε ο χρήστης (hairpin ή mature). Επιστρέφει έναν πίνακα με σύνθετα στοιχεία καθένα από τα οποία περιέχει πληροφορίες για κάθε μια δημοσίευση που προέκυψε από την αναζήτηση στην βάση.

TooltipManager Component

Το component αυτό περιέχει περιγραφές για διάφορα εργαλεία κάθε εφαρμογής της ιστοσελίδας DIANA. Αρχικοποιείται στο configuration αρχείο main.php και στη συνέχεια τα στοιχεία του χρησιμοποιούνται για την προβολή αυτών των περιγραφών στον χρήστη.

MirpubHeader widget

Η κλάση αυτή δημιουργεί τη φόρμα αναζήτησης στην εφαρμογή mirPub και προβάλλεται στη διεπαφή μέσω του php αρχείου diana_header. Χρησιμοποιείται τόσο από την αρχική σελίδα της εφαρμογής (index.php) όσο και από τη σελίδα όπου εμφανίζονται τα αποτελέσματα (prefix_results.php).

prefix_results View

Μέσω του συγκεκριμένου αρχείου προβάλλονται στον χρήστη τα αποτελέσματα της αναζήτησης που έκανε. Καλείται από τη μέθοδο actionResults του Controller.

results_page View

Παράγει την τελική σελίδα που εμφανίζεται στον χρήστη μετά την αναζήτηση. Επομένως περιέχει και τα αποτελέσματα της prefix_results.php. Καλείται από τη μέθοδο actionGetResults του Controller.

5.2 Ανάλυση απαιτήσεων συστήματος

Η υλοποίηση του μηχανισμού κατάταξης μπορεί να χωριστεί σε τρία διακριτά βήματα. Πρέπει αρχικά να παρέχεται γραφικά στον χρήστη η δυνατότητα επιλογής της μεθόδου κατάταξης και στη συνέχεια τα αποτελέσματα να ταξινομούνται με βάση την επιλογή αυτή. Τέλος με την εισαγωγή νέων δημοσιεύσεων πρέπει να επανυπολογίζονται οι τιμές των μετρικών για όλες τις δημοσιεύσεις.

5.2.1 Εμφάνιση λίστας για δυνατότητα επιλογής αλγόριθμου κατάταξης

Δίπλα στη φόρμα αναζήτησης θα πρέπει να εμφανίζεται στον χρήστη μια drop-down λίστα με τις διάφορες τεχνικές με τις οποίες μπορούν να ταξινομηθούν τα αποτελέσματα της αναζήτησης. Οι τεχνικές αυτές είναι:

- Χρονολογική κατάταξη.
- Κατάταξη με βάση τον αλγόριθμο Pagerank.
- Κατάταξη με βάση τον αλγόριθμο FutureRankLike.
- Κατάταξη με βάση τον αλγόριθμο AdvRecPubs-RankWithIF (για αισθητικούς λόγους, στην εφαρμογή εμφανίζεται μόνο το πρώτο συνθετικό του ονόματος του αλγόριθμου).

Θα πρέπει επίσης να παρουσιάζεται μια μικρή περιγραφή αυτών των αλγορίθμων όταν ο δείκτης του ποντικιού του χρήστη περνάει πάνω από το όνομα κάθε στοιχείου της λίστας.

5.2.2 Κατάταξη αποτελεσμάτων αναζήτησης με βάση τον επιλεγμένο αλγόριθμο

Ανάλογα με την επιλογή του χρήστη από τη λίστα της διεπαφής, τα αποτελέσματα θα πρέπει να ταξινομούνται και να επιστρέφονται με την επιθυμητή κατάταξη. Αυτό σημαίνει ότι οι τιμές των μετρικών για κάθε δημοσίευση θα πρέπει να βρίσκονται αποθηκευμένες στη βάση. Για όλες λοιπόν τις δημοσιεύσεις που πρέπει να παρουσιαστούν στα αποτελέσματα της αναζήτησης, μαζί με τα υπόλοιπα στοιχεία τους (τίτλος, χρονολογία δημοσίευσης κ.τ.λ.) θα πρέπει να αντλούνται από τη βάση και οι τιμές των μετρικών έτσι ώστε να πραγματοποιείται η επιθυμητή κατάταξη.

5.2.3 Επανυπολογισμός μετρικών μετά την εισαγωγή νέων δημοσιεύσεων στο σύστημα

Όταν εισάγονται νέες δημοσιεύσεις στη βάση, θα πρέπει να επανυπολογίζονται οι μετρικές τόσο για τις νέες δημοσιεύσεις οι οποίες δεν έχουν ακόμα τιμή για την κάθε μετρική, όσο και για τις παλιές, εξαιτίας της εισαγωγής νέων ακμών στον γράφο που χρησιμοποιούν σαν είσοδο οι μέθοδοι κατάταξης.

5.3 Πλατφόρμες και προγραμματιστικά εργαλεία

Στη συνέχεια περιγράφονται οι πλατφόρμες και τα προγραμματιστικά εργαλεία που χρησιμοποιήθηκαν για την ανάπτυξη της ζητούμενης λειτουργικότητας.

5.3.1 Apache HTTP server

Ο Apache Web Server είναι ένας εξυπηρετητής (server) του παγκόσμιου Ιστού (Web). Είναι συμβατός με διάφορα λειτουργικά συστήματα όπως Linux, Unix, Microsoft Windows, GNU, FreeBSD, Solaris, Novell NetWare, Mac OS X, OS/2, TPF. Ο ρόλος του Apache είναι να αναμένει αιτήσεις από διάφορα προγράμματα – χρήστες (clients) όπως ένας browser ενός χρήστη και στη συνέχεια να εξυπηρετεί αυτές τις αιτήσεις "σερβίροντας" τις σελίδες που ζητούν είτε απευθείας μέσω μιας ηλεκτρονικής διεύθυνσης (URL), είτε μέσω ενός συνδέσμου (link). Ο τρόπος με τον οποίο ο Apache εξυπηρετεί αυτές τις αιτήσεις, είναι σύμφωνος με τα πρότυπα που ορίζει το πρωτόκολλο HTTP (Hypertext Transfer Protocol).

Ο Apache HTTP αναπτύσσεται από την "Κοινότητα Ανοιχτού Λογισμικού" και ο έλεγχος, η υποστήριξη, και η διάθεση του προγράμματος γίνεται από το Apache Software Foundation. Το πρόγραμμα αυτό διατίθεται δωρεάν και μπορούν να γίνουν ελεύθερα από το χρήστη προσθήκες και τροποποιήσεις στον κώδικα του.

Ο Apache διαθέτει ποικιλία χαρακτηριστικών και μπορεί να υποστηρίξει μια μεγάλη γκάμα εφαρμογών. Ένα από τα βασικότερα χαρακτηριστικά του, είναι ότι μπορεί να προσαρμόσει επάνω του πολλές προσθήκες προγραμμάτων (modules), τα οποία με τη σειρά τους παρέχουν διαφορετικές λειτουργίες. Μερικά από τα πιο γνωστά modules του Apache HTTP είναι τα modules πιστοποίησης, όπως για παράδειγμα τα mod_access, mod_auth, mod_digest κ.λ.π. Πραγματοποιεί επίσης ανακατευθύνσεις διευθύνσεων (URL rewrites) μέσω του mod_rewrite, καταγραφές συνδέσεων μέσω του mod_log_config, συμπίεση αρχείων μέσω του mod_gzip. Διαθέτει πολλά ακόμη modules τα οποία διατίθενται είτε απο το Apache Software Foundation, είτε από τρίτες εταιρίες λογισμικού. Ο Apache HTTP υποστηρίζει επίσης αρκετές διάσημες εφαρμογές και γλώσσες προγραμματισμού όπως MySQL, PHP, Perl, Python κ.λ.π.

Αυτά είναι μερικά από τα χαρακτηριστικά και τις λειτουργίες του που κάνουν τον Apache τον πιο δημοφιλή Web Server από το 1996 έως τις μέρες μας. Περισσότερο από το 50% των ιστοχώρων του παγκόσμιου ιστού, χρησιμοποιεί τον Apache ως εξυπηρετητή.

5.3.2 mySQL

Η mySQL είναι ένα σύστημα διαχείρισης σχεσιακών βάσεων ανοικτού κώδικα, που χρησιμοποιεί τη Structured Query Language (SQL), την πιο γνωστή γλώσσα για την προσθήκη, πρόσβαση και επεξεργασία εγγραφών σε μία Βάση Δεδομένων. Η mySQL είναι γνωστή κυρίως για την ταχύτητα, την αξιοπιστία, και την ευελιξία που παρέχει.

Κάποια από τα κύρια χαρακτηριστικά της είναι τα εξής:

- Είναι υλοποιημένη σε C και C++
- Αποτελεί σύστημα πελάτη-εξυπηρετητή, όπου εξυπηρετητής είναι η βάση και πελάτες διάφορες εφαρμογές, εγκατεστημένες στο ίδιο ή διαφορετικό σύστημα που επικοινωνούν με τη βάση εκτελώντας αιτήματα σε αυτήν και αλλάζοντας τις εγγραφές της.
- Υπάρχει ένας μεγάλος αριθμός APIs (application programming interfaces) και βιβλιοθηκών για την ανάπτυξη mySQL εφαρμογών. Τα συστήματα-πελάτες μπορούν να είναι υλοποιημένα σε C, C++, Java, Perl, PHP και Python.
- Εκτός από τα συστήματα-πελάτες, και το σύστημα-εξυπηρετητή, μπορεί να τρέξει σε διάφορα λειτουργικά συστήματα όπως Apple Macintosh OS, Linux, Microsoft Windows και Sun Solaris.
- Υποστηρίζει όλα τα σύνολα χαρακτήρων.
- Παρέχεται η δυνατότητα για αποθήκευση και χρήση διαδικασιών (procedures), έτσι ώστε να απλοποιούνται κάποια βήματα όπως η εισαγωγή ή η διαγραφή εγγραφών στη βάση. Για τα συστήματα-πελάτες δίνεται έτσι η δυνατότητα να μην τροποποιούν απευθείας τους πίνακες της βάσης. Διασφαλίζεται επίσης η αποτελεσματική διαχείριση μεγάλων βάσεων δεδομένων.
- Είναι σχεδιασμένη για πολυνηματική λειτουργία καθώς και για χρήση πολλών επεξεργαστών.
- Παρέχεται η δυνατότητα χρήσης transactions. Όσον αφορά τις βάσεις δεδομένων ένα transaction είναι μία σειρά ενεργειών που εκτελούνται στη βάση σαν ενιαίο σύνολο και το σύστημα εξασφαλίζει ότι είτε θα εκτελεστούν όλες είτε καμία απο αυτές. Το προεπιλεγμένο πρότυπο πινάκων που ονομάζεται myISAM δεν υποστηρίζει transactions. Παρ' όλα αυτά, υπάρχουν πρότυπα που τα υποστηρίζουν με γνωστότερο το InnoDB

5.3.3 PHP

Η γλώσσα προγραμματισμού PHP είναι μια ευρέως χρησιμοποιούμενη γλώσσα γενικού σκοπού και ανοιχτού κώδικα, που προορίζεται κυρίως για την ανάπτυξη δικτυακών εφαρμογών. Μπορεί να εγκατασταθεί σχεδόν σε όλα τα λειτουργικά συστήματα όπως Windows, Linux, Mac OS X, Risc OS κ.λ.π. και υποστηρίζεται από τους περισσότερους εξυπηρετητές ιστοσελίδων όπως ο Apache ή ο IIS. Η PHP μπορεί να λειτουργήσει είτε ως εγκατεστημένη

μονάδα (module) στον εξυπηρετητή ιστοσελίδων είτε μέσω ενός επεξεργαστή CGI σεναρίων. Μπορεί να χρησιμοποιηθεί για εκτέλεση σεναρίων (scripts) από την πλευρά του απομακρυσμένου εξυπηρετητή ιστοσελίδων όπως γίνεται και με τα σενάρια CGI. Επίσης η PHP μπορεί να χρησιμοποιηθεί για είσοδο/έξοδο δεδομένων από τον χρήστη ή για τη δυναμική δημιουργία σελίδων.

Σενάρια PHP σε απομακρυσμένο εξυπηρετητή ιστοσελίδων

Αυτή είναι η κύρια χρήση της γλώσσας PHP. Η γλώσσα PHP βρίσκεται εγκαταστημένη είτε ως module στον εξυπηρετητή ιστοσελίδων είτε εκτελείται μέσω ενός CGI σεναρίου και χρησιμοποιείται δια μέσου ενός φυλλομετρητή από τον υπολογιστή του πελάτη-χρήστη. Ο κώδικας της PHP είναι κώδικας που εκτελείται στον εξυπηρετητή, παράγει κώδικα HTML και στη συνέχεια στέλνεται στον πελάτη-χρήστη. Ο πελάτης λαμβάνει επομένως τα αποτελέσματα της εκτέλεσης του κώδικα ενός αρχείου PHP.

Χρήση της PHP σε επίπεδο γραμμής εντολών (command line)

Ένα σενάριο PHP μπορεί να εκτελεστεί μέσω του διερμηνέα της τοπικά στον υπολογιστή χωρίς να χρειάζεται να μεσολαβήσει ένας εξυπηρετητής ιστοσελίδων.

Εφαρμογές με τη γλώσσα PHP

Αν και δε συνηθίζεται, μπορούν να φτιαχτούν προγράμματα με γραφικό περιβάλλον (π.χ. χρησιμοποιώντας το PHP-GTK) που να τρέχουν κατευθείαν στον υπολογιστή πελάτη τα οποία είναι ανεξάρτητα πλατφόρμας.

Το συντακτικό της PHP είναι βασισμένο στη σύνταξη της γλώσσας C, Java και Perl και είναι εύκολη στην εκμάθηση.

5.3.3.1 Το Yii framework

Το Yii είναι ένα PHP framework ιδανικό για την ανάπτυξη δικτυακών εφαρμογών. Τα σημαντικότερα χαρακτηριστικά του είναι:

MVC

Είναι βασισμένη στο σχεδιαστικό πρότυπο MVC (model-view-controller) το οποίο χρησιμοποιείται ευρέως στον δικτυακό προγραμματισμό και σκοπός του είναι ο διαχωρισμός της λογικής της εφαρμογής από τη διεπαφή έτσι ώστε ο προγραμματιστής να μεταβάλλει εύκολα το ένα μέρος χωρίς να επηρεάζεται το άλλο. Στο model αποθηκεύονται τα δεδομένα και η λογική της εφαρμογής, το view περιέχει τα στοιχεία της διεπαφής και ο controller ελέγχει την επικοινωνία μεταξύ model και view.

- Αντικείμενο μιας κλάσης controller δημιουργείται από την εφαρμογή όταν υπάρχει η αντίστοιχη αίτηση. Όταν τρέχει, εκτελεί τη ζητούμενη ενέργεια φέρνοντας το κατάλληλο model και απεικονίζοντάς το με το κατάλληλο view. Η απλούστερη δομή του είναι η ενέργεια (action) που είναι απλώς μία μέθοδος της κλάσης, της οποίας το όνομα ξεκινάει με τη λέξη action. Ο κάθε controller έχει μια προεπιλεγμένη

ενέργεια που ονομάζεται index και χρησιμοποιείται όταν η αίτηση δε διευκρινίζει ποια ακριβώς ενέργεια του controller πρέπει να εκτελεστεί. Στο Yii framework υπάρχει και ένας αρχικός controller που ονομάζεται application, ο οποίος λαμβάνει την αίτηση του χρήστη και τη στέλνει στον κατάλληλο controller.

- Στο Yii framework υλοποιούνται δύο είδη models: Το form model και το active record. Το form model χρησιμοποιείται για την αποθήκευση δεδομένων που εισάγονται και η χρήση τους είναι συνήθως προσωρινή. Το Active Record (AR) είναι ένα σχεδιαστικό πρότυπο που χρησιμοποιείται για την επικοινωνία με μία βάση δεδομένων και την επεξεργασία των εγγραφών της με αντικειμενοστρεφή τρόπο. Κάθε αντικείμενο AR αντιπροσωπεύει μια εγγραφή ενός πίνακα της βάσης και οι ιδιότητες του αντικειμένου αντιπροσωπεύουν τα πεδία της εγγραφής αυτής.
- Κάθε κλάση view είναι ένα PHP script που αποτελείται κυρίως από στοιχεία που αφορούν στη διεπαφή με το χρήστη.
- Υπάρχουν επίσης και οι κλάσεις widget που το περιεχόμενό τους, όπως και των views, έχει να κάνει κυρίως με τη διεπαφή. Ενσωματώνονται συνήθως σε ένα view και αποτελούν ένα αυτοτελές κομμάτι της διεπαφής. Η χρήση τους δίνει τη δυνατότητα επαναχρησιμοποίησης του κώδικά τους από πολλά views.

DAO/ActiveRecord

Το framework Yii υποστηρίζει τον προγραμματισμό με βάσεις δεδομένων. Τα DAO (Data Access Objects) αντικείμενα του Yii παρέχουν τη δυνατότητα επικοινωνίας με διαφορετικά συστήματα βάσεων δεδομένων (Database Management Systems DBMS) χρησιμοποιώντας μία και μόνο διεπαφή. Αναπαριστώντας έναν πίνακα με τη μορφή μιας κλάσης και μία εγγραφή του με τη μορφή ενός αντικειμένου της κλάσης αυτής, αποφεύγεται η εγγραφή SQL κώδικα για την επεξεργασία των εγγραφών της βάσης.

Caching

Το framework Yii παρέχει τη δυνατότητα χρήσης προσωρινής μνήμης για την ταχεία προσπέλαση δεδομένων που χρησιμοποιούνται συχνά.

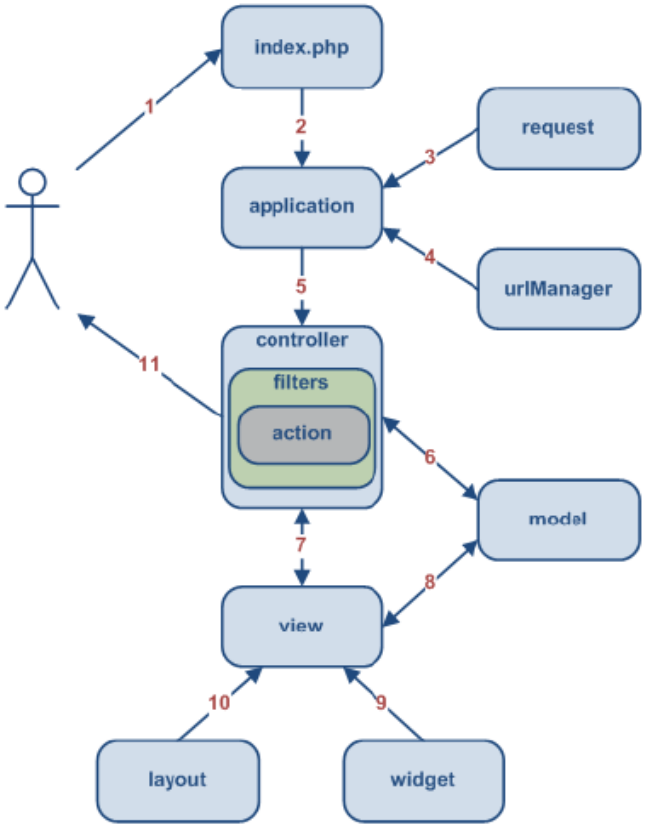
Παρακάτω περιγράφουμε μία συνήθη ροή εργασιών μιας εφαρμογής υλοποιημένης με το Yii framework:

1. Ο χρήστης πραγματοποιεί μια αίτηση στη διεύθυνση που βρίσκεται η εφαρμογή και ο εξυπηρετητής εκτελεί το script index.php που βρίσκεται στη διεύθυνση αυτή.

2. Το script αυτό αρχικοποιεί και εκτελεί τον αρχικό controller application.

3. Ο controller application λαμβάνει την κλήση του χρήστη μέσω ενός συστατικού που λέγεται request.

4. Ο controller application προσδιορίζει ποιος controller πρέπει να δημιουργηθεί και ποια ενέργειά του (action) πρέπει να εκτελεστεί μέσω ενός συστατικού που ονομάζεται urlManager.



Σχήμα 5.1: Συνήθης ροή εργασιών μιας εφαρμογής υλοποιημένης με το Yii framework

5. Ο controller application δημιουργεί ένα αντικείμενο της αντίστοιχης controller κλάσης. Το αντικείμενο αυτό εκτελεί τη ζητούμενη ενέργεια (action) αν το επιτρέπουν τα φίλτρα που έχουν εφαρμοστεί στο action αυτό.

6. Το action διαβάζει το αντικείμενο μιας κλάσης model από τη βάση δεδομένων.

7. Ενεργοποιείται το κατάλληλο view.

8. Το view αυτό διαβάζει και εμφανίζει τα δεδομένα από το αντικείμενο model.

9. Εκτελούνται τα widgets που υπάρχουν στο παραπάνω view αντικείμενο.

10. Όλες οι ενέργειες που πραγματοποιήθηκαν σε επίπεδο view ενσωματώνονται στη διεπαφή.

11. Τα αποτελέσματα εμφανίζονται στον χρήστη.

5.3.4 XAMPP

Το XAMPP αποτελεί μία πλατφόρμα ελεύθερου λογισμικού, η οποία περιέχει τον εξυπηρετητή Apache, τη βάση δεδομένων mySQL και διερμηνέα (interpreter) για κώδικα γραμμένο σε γλώσσες προγραμματισμού PHP και Perl.

Προορίζεται ως εργαλείο ανάπτυξης και δοκιμής ιστοσελίδων τοπικά στον υπολογιστή χωρίς να είναι απαραίτητη η σύνδεση στο Διαδίκτυο. Το XAMPP υποστηρίζει τη δημιουργία και διαχείριση βάσεων δεδομένων τύπου mySQL και SQLite.

5.4 Περιγραφή - επίδειξη λειτουργικότητας

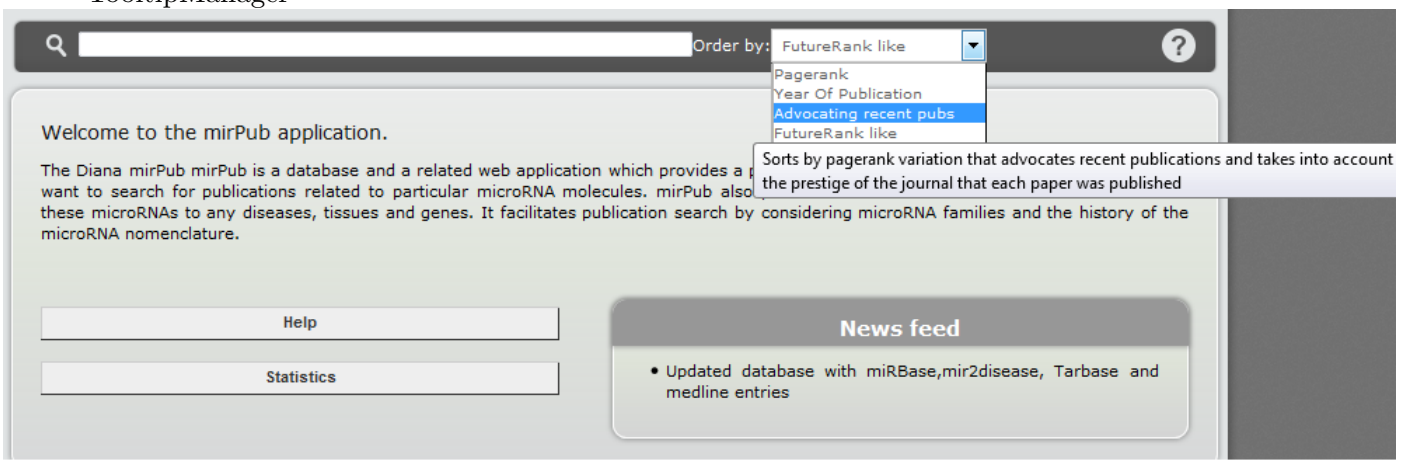
Λαμβάνοντας υπόψη τις απαιτήσεις που αναλύθηκαν παραπάνω, υλοποιήθηκε η ζητούμενη λειτουργικότητα. Χρησιμοποιήθηκε το εργαλείο XAMPP έτσι ώστε να εγκατασταθεί η εφαρμογή τοπικά και να πραγματοποιούνται εύκολα οι οποιεσδήποτε αλλαγές. Στην ουσία, τροποποιήθηκαν οι κλάσεις της εφαρμογής και ο πίνακας της βάσης της που αναφέρθηκαν παραπάνω.

5.4.1 Εμφάνιση λίστας για δυνατότητα επιλογής αλγόριθμου κατάταξης

Το widget MirpubHeader χρησιμοποιείται από την εφαρμογή για την προβολή της φόρμας αναζήτησης. Το widget αυτό καλεί με τη σειρά του το PHP script τύπου view, diana_header, στο οποίο πραγματοποιήθηκαν οι αλλαγές. Για τη δημιουργία της drop-down λίστας χρησιμοποιήθηκε η μέθοδος dropDownList της κλάσης CHtml του Yii framework, η οποία έχει σαν ορίσματα το όνομα, την προεπιλεγμένη τιμή, δεδομένα για την προβολή των διαθέσιμων επιλογών και επιπρόσθετα HTML attributes της λίστας και σαν έξοδο παράγει σε HTML κώδικα τη ζητούμενη λίστα.

Σχήμα 5.2: Η φόρμα αναζήτησης μετά την προσθήκη της drop-down λίστας

Για την παρουσίαση μιας μικρής περιγραφής κάθε τεχνικής κατάταξης στον χρήστη, προστέθηκαν στο τελευταίο όρισμα της μεθόδου dropDownList τα αντίστοιχα HTML attributes για κάθε διαθέσιμη επιλογή της λίστας. Οι περιγραφές αυτές προστέθηκαν στο Component TooltipManager



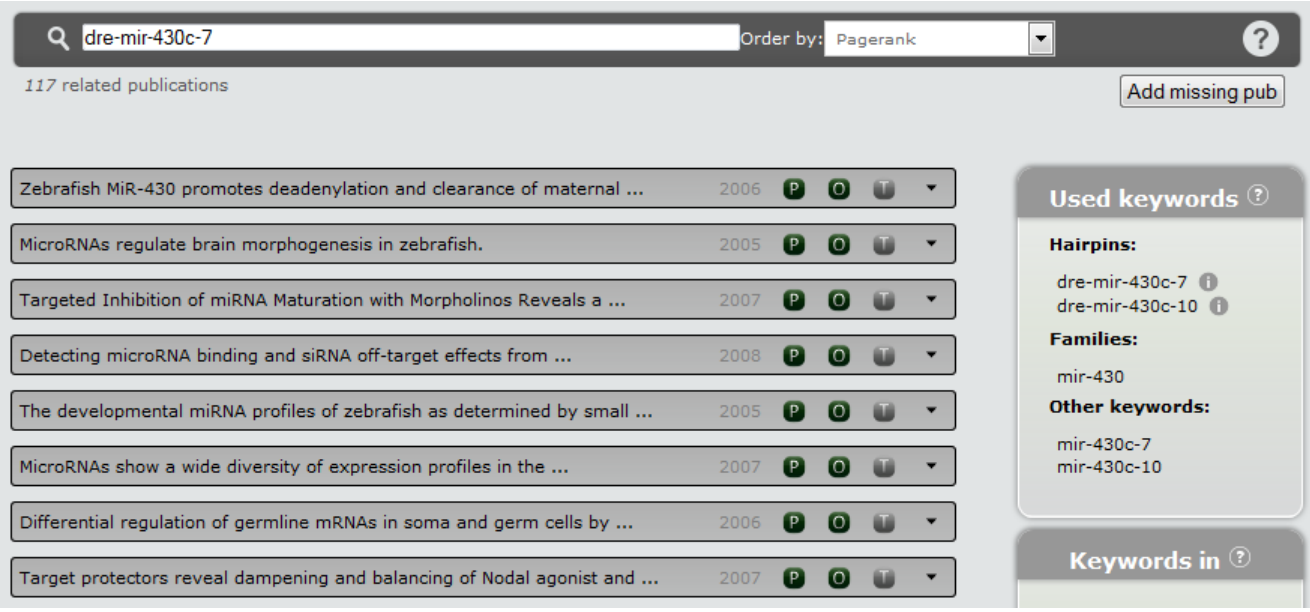
Σχήμα 5.3: Περιγραφή του αλγόριθμου κατάταξης AdvRecPubs-RankWithIF

5.4.2 Ταξινόμηση αποτελεσμάτων αναζήτησης με βάση τον επιλεγμένο αλγόριθμο

1. Στον Πίνακα diana_paper της βάσης προστέθηκαν 3 στήλες τύπου double, μία για τον κάθε αλγόριθμο κατάταξης.

2. Χρησιμοποιήθηκε το python script που περιγράφεται στην Ενότητα 5.4.3, που παίρνει σαν είσοδο το όνομα ενός αρχείου το οποίο περιέχει τα pubmed ids (αναγνωριστικά των δημοσιεύσεων στη βάση PubMed) των νέων δημοσιεύσεων και ανανεώνει τις τιμές των μετρικών κάθε δημοσίευσης στη βάση.

3. Τροποποιήθηκε η μέθοδος get_citations_by_type_array της model κλάσης CitationFetcher έτσι ώστε ο πίνακας που επιστρέφει να περιέχει και τις τιμές των μετρικών κάθε δημοσίευσης.



Σχήμα 5.4: Μέρος των αποτελεσμάτων αναζήτησης με λέξη - κλειδί dre-mir-430c-7, ταξινομημένα με τη χρήση του αλγόριθμου pagerank

4. Το script τύπου view το οποίο είναι υπεύθυνο για την κατάταξη και την εμφάνιση των αποτελεσμάτων στη διεπαφή είναι το prefix_results. Έπρεπε λοιπόν η επιλογή του χρήστη από την drop-down λίστα να αποθηκεύεται σε μια μεταβλητή και να μεταφέρεται στο script αυτό, έτσι ώστε η κατάταξη να γίνεται με βάση την τιμή της μεταβλητής και όχι απαραίτητα με βάση τη χρονολογία δημοσίευσης, όπως γινόταν μέχρι τώρα. Γι' αυτό, η μεταβλητή ενσωματώθηκε σαν HTTP GET παράμετρος στις μεθόδους actionResults και actionGetResults του MirpubController καθώς και στο script results_page.php.

5.4.3 Επανυπολογισμός μετρικών μετά την εισαγωγή νέων δημοσιεύσεων στο σύστημα

Τα python scripts που υλοποιούν τους επιλεγμένους αλγόριθμους κατάταξης (Κεφάλαιο 3) τροποποιήθηκαν ώστε να δημιουργηθεί ένα script το οποίο πρόκειται να ενσωματωθεί στο σύστημα. Το σύστημα αυτό, έχει σαν είσοδο το όνομα ενός αρχείου όπου καταγράφονται τα pubmed ids των νέων δημοσιεύσεων της βάσης, ανανεώνει τις τιμές των μετρικών των δημοσιεύσεων στη βάση και θα τρέχει κάθε φορά που εισάγονται νέες δημοσιεύσεις. Για την επικοινωνία με τη βάση της εφαρμογής χρησιμοποιήθηκε η βιβλιοθήκη της Python mysql.connector.

Με την εκτέλεση του script πραγματοποιούνται τα παρακάτω βήματα:

1. Ο γράφος που δημιουργείται από τα άρθρα και τις παραπομπές τους, εμπλουτίζεται με τους νέους κόμβους και τις ακμές τους που συλλέγονται από τα προγράμματα Efetch και Elink. Ο γράφος αυτός είναι αποθηκευμένος σαν binary μεταβλητή τοπικά, ώστε να μην ανακτώνται κάθε φορά οι παραπομπές όλων των δημοσιεύσεων μέσω των Efetch και Elink παρά μόνο των νέων δημοσιεύσεων.

2. Συλλέγονται από τη βάση οι χρονολογίες και τα περιοδικά δημοσίευσης των νέων άρθρων και ανανεώνονται οι binary μεταβλητές που είναι αποθηκευμένες τοπικά τύπου dictionary με keys τα pubmed ids των άρθρων και values τις ημερομηνίες και τα περιοδικά δημοσίευσής τους.

3. Υπολογίζονται οι τιμές του αλγόριθμου PageRank με τη χρήση του παραπάνω γράφου.

4. Υπολογίζονται οι τιμές του αλγόριθμου FutureRankLike με τη χρήση του παραπάνω γράφου και των μεταβλητών με τις χρονολογίες και τα περιοδικά δημοσίευσης των άρθρων.

5. Υπολογίζονται οι τιμές του αλγόριθμου AdvRecPubs-RankWithIF με τη χρήση του παραπάνω γράφου και των μεταβλητών με τις χρονολογίες και τα περιοδικά δημοσίευσης των άρθρων.

6. Τα αποτελέσματα εισάγονται στη βάση. Σε περίπτωση που έστω και ένα από τα αιτήματα στη βάση αποτύχουν, ανακτάται η προηγούμενη εικόνα του πίνακα.

Κεφάλαιο 6

Επίλογος

Ανακεφαλαιώνοντας, παρουσιάζεται το αντικείμενο της διπλωματικής εργασίας και επισημαίνονται οι χρήσεις των αποτελεσμάτων της. Τέλος, γίνεται αναφορά σε πιθανές μελλοντικές επεκτάσεις.

6.1 Σύνοψη

Η παρούσα διπλωματική εργασία είχε ως αντικείμενο την υλοποίηση μηχανισμού κατάταξης των δημοσιεύσεων που είναι καταχωρημένες στην εφαρμογή Diana Mirpub, ο οποίος θα βασίζεται στη σημαντικότητα των εργασιών αυτών. Στη βάση αυτή, μελετήθηκε ο αλγόριθμος PageRank, ο οποίος αποτελεί μέρος του μηχανισμού κατάταξης των ιστοσελίδων στο Διαδίκτυο, καθώς και σχετικές εργασίες που προτείνουν εναλλακτικούς μηχανισμούς κατάταξης, λαμβάνοντας υπόψη τις διαφορές του παγκόσμιου ιστού με το δίκτυο των δημοσιεύσεων. Στη συνέχεια αναπτύχθηκαν προγράμματα που υλοποιούν τους παραπάνω αλγόριθμους, τα οποία εκτελέστηκαν με είσοδο τα δεδομένα της βάσης της εφαρμογής. Οι μηχανισμοί με τα καλύτερα αποτελέσματα ενσωματώθηκαν στη διεπαφή της εφαρμογής. Οι μηχανισμοί κρίθηκαν με βάση τη μέθοδο αξιολόγισης που αναπτύχθηκε, ή οποία βασίζεται στην εφαρμογή της μετρικής Spearman στα αποτελέσματα των υλοποιημένων αλγορίθμων. Τελικό προϊόν της εργασίας αποτελεί η νέα λειτουργικότητα που ενσωματώθηκε στο λογισμικό της εφαρμογής Diana Mirpub. Η λειτουργικότητα αυτή δίνει τη δυνατότητα στον χρήστη να επιλέξει τον τρόπο που πρόκειται να καταταχθούν τα αποτελέσματα μιας αναζήτησης.

6.2 Μελλοντικές επεκτάσεις

Η λειτουργικότητα που υλοποιήθηκε μπορεί να επεκταθεί ώστε να διευκολύνει ακόμα περισσότερο τον χρήστη στην κατεύθυνση εύρεσης των πιο χρήσιμων γι' αυτόν εργασιών. Στη συνέχεια παρατίθενται κάποιες προτάσεις μελλοντικών επεκτάσεων.

Εμπλουτισμός του δικτύου δημοσιεύσεων

Το σύνολο των δεδομένων που χρησιμοποιείται σαν είσοδος των αλγόριθμων κατάταξης αντλείται από τις δημοσιεύσεις που είναι καταχωρημένες στη βάση της εφαρμογής. Τόσο

οι κόμβοι (δημοσιεύσεις), όσο και οι ακμές (παραπομπές) του γράφου που δημιουργείται για την εφαρμογή των μηχανισμών κατάταξης, περιορίζονται στα δεδομένα αυτά. Όπως είδαμε στο Κεφάλαιο 4 των αποτελεσμάτων, το μεγαλύτερο πλήθος των παραπομπών που συναντάται στη βιβλιογραφία των εργασιών αγνοείται και ο γράφος που δημιουργείται είναι αρκετά αραιός, γεγονός που επηρεάζει τα αποτελέσματα των σκορ κατάταξης τόσο των εργασιών, όσο και των περιοδικών. Αν θεωρήσουμε τον γράφο που δημιουργείται από όλες τις δημοσιεύσεις της βάσης PubMed, το ζητούμενο είναι για τον υπολογισμό των σκορ κατάταξης να χρησιμοποιήσουμε έναν υπογράφο που να περιορίζεται όσο το δυνατόν περισσότερο στο συγκεκριμένο επιστημονικό αντικείμενο και οι περισσότερες ακμές των κόμβων του γράφου που αποτελούν και κόμβους του υπογράφου, να κατευθύνονται σε κόμβους που ανήκουν επίσης στον υπογράφο. Είναι πιθανόν, στο μέλλον προσθέτοντας ακόμα περισσότερες δημοσιεύσεις στη βάση της εφαρμογής, αυτό να επιτευχθεί. Παρ' όλα αυτά, σαν πρώτη προσπάθεια, θα μπορούσαν να εισαχθούν στον γράφο σαν κόμβοι όλες οι εργασίες στις οποίες παραπέμπουν οι υπάρχουσες εργασίες του γράφου καθώς και οι ακμές αυτών των εργασιών που παραπέμπουν σε υπάρχουσες εργασίες του γράφου.

Υλοποίηση προφίλ χρηστών

Ισχύει ότι οι προτιμήσεις των χρηστών μιας μηχανής αναζήτησης ποικίλουν. Με τον ίδιο τρόπο διαφοροποιείται και το πεδίο ενδιαφέροντος των ερευνητών που χρησιμοποιούν την εφαρμογή Diana Mirpub, με σκοπό να αντλήσουν πληροφορίες σχετικές με κάποιο micro-RNA. Για παράδειγμα, ένας από αυτούς μπορεί να αναζητά το σύνολο των ασθενειών που σχετίζονται με ένα micro-RNA, ενώ κάποιος άλλος τη δομή του συγκεκριμένου μορίου. Προτείνεται λοιπόν η υλοποίηση προφίλ χρηστών για καταγραφή των προτιμήσεών τους. Οι προτιμήσεις αυτές μπορούν να αφορούν σε συγκεκριμένο επιστημονικό πεδίο ή στο σύνολο περιοδικών, στο οποίο οι εργασίες που επιστρέφονται αναμένεται να είναι δημοσιευμένες. Μπορούν επίσης να καταχωρούνται απ' ευθείας από τον χρήστη ή να αντλούνται από το ιστορικό του.

Κατάταξη με βάση τις κλασσικές μετρικές ανάκτησης πληροφορίας

Μια σημαντική εργασία δεν αποτελεί απαραίτητα το καλύτερο δυνατό αποτέλεσμα της αναζήτησης με βάση κάποιο ερώτημα. Ο χρήστης πρώτα απ' όλα επιθυμεί τα αποτελέσματα να είναι όσο το δυνατόν πιο σχετικά με το ερώτημα που έθεσε. Θα μπορούσαμε να πούμε ότι η σχετικότητα μιας εργασίας με έναν όρο, είναι ανάλογη με την εμφάνιση του όρου αυτού πρώτα από όλα στον τίτλο της και στη συνέχεια με το πλήθος των εμφανίσεων του όρου στην περίληψη και το κείμενό της. Σημαντικό ρόλο παίζει επίσης και το αντίστοιχο πλήθος εμφανίσεων των υπόλοιπων όρων με βάση τους οποίους θα μπορούσε να πραγματοποιηθεί μια αναζήτηση. Όπως αναφέρεται στην Ενότητα 2.1 της εισαγωγής, τα συστήματα ανάκτησης διαθέτουν τέτοιο μηχανισμό ταξινόμησης των αποτελεσμάτων, χρησιμοποιώντας το ευρετήριο που έχει δημιουργηθεί κατά τη διαδικασία της δεικτοδότησης (indexing). Αυτή η διαδικασία θα μπορούσε να υιοθετηθεί και στην περίπτωσή μας και η τελική κατάταξη να αποτελεί συνδυασμό των κατατάξεων που

προκύπτουν με βάση τη σχετικότητα των αποτελεσμάτων και το κύρος των εργασιών.

Βιβλιογραφία

[1] https://secure.php.net/.

[2] Apache Http Server Project, http://httpd.apache.org/.

[3] dblp: computer science bibliography, http://dblp.uni-trier.de/.

[4] Diana Mirpub, http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=site/index.

[5] EIGENFACTOR.org, http://www.eigenfactor.org.

[6] Eric Sayers, 'The E-utilities In-Depth: Parameters, Syntax and More.' http://www.ncbi.nlm.nih.gov/books/NBK25499/, May 2009.

[7] Erjia Yan, Ying Ding, 'Weighted citation: An indicator of an article's prestige.' School of Library and Information Science, Indiana University, Bloomington, USA, 2010.

[8] Hassan Sayyadi, Lise Getoory, 'FutureRank: Ranking Scientific Articles by Predicting their Future PageRank.' SIAM International Conference on Data Mining, 2009.

[9] Ilias Kanellos, Vasiliki Vlachokyriakou, Thanasis Vergoulis, Georgios Georgakillas, Yannis Vasileiou, Artemis Hatzigeorgiou and Theodore Dalamagas., *MirPub v2: Towards Ranking and Refining miRNA Publication Search Results.* TPDL 2015.

[10] Impact Factor Search, http://www.journal-database.com.

[11] Introduction to MySQL Connector/Python, http://dev.mysql.com/doc/connector-python/en/connector-python-introduction.html.

[12] J West, CT Bergstrom, 'Pseudocode for calculating Eigenfactor TM Score and Article Influence TM Score using data from Thomson-Reuters Journal Citations Reports.' 2008.

[13] JASIST: Journal of the Association for Information Science and Technology, https://www.asis.org/jasist.html.

[14] Judit Bar-Ilan, 'Comparing rankings of search results on the Web.' 2005.

[15] L Page, S Brin, R Motwani, T Winograd, 'The PageRank Citation Ranking: Brinking Order to the Web.' 1999.

[16] MeSH, https://www.nlm.nih.gov/pubs/factsheets/mesh.html.

[17] Michael Buckland, Fredric Gey, 'The Relationship between Recall and Precision, page 15 (Precision versus Recall),' School of Library and Information Studies, University of California, Berkeley, 1994.

[18] National Center for Biotechnology Information (US), 'Entrez Programming Utilities Help - NCBI help manual.' http://www.ncbi.nlm.nih.gov/books/NBK25501/, 2010, Chapter: The E-utilities In-Depth: Parameters, Syntax and More.

[19] NCBI Bookshelf, http://www.ncbi.nlm.nih.gov/books/.

[20] NetworkX High-productivity software for complex networks, https://networkx.github.io/.

[21] Pagerank, https://en.wikipedia.org/wiki/PageRank.

[22] P.Chena, H.Xie, S.Maslovc, S.Rednera, 'Finding scientific gems with Google's PageRank algorithm.'

[23] Pubmed, http://www.ncbi.nlm.nih.gov/pubmed.

[24] Python, https://www.python.org/.

[25] Qiang Xue and Xiang Wei Zhuo, *The Defnitive Guide to Yii 1.1.* Yii Software LLC, 2008-2010 (c).

[26] Ronald Fagin, Ravi Kumar, D. Sivakumar, 'Comparing top k lists.' IBM Almaden Research Center, 2003.

[27] Ronald Fagin, Ravi Kumar, Mohammad Mahdian, D. Sivakumar, Erik Vee, 'Comparing and Aggregating Rankings with Ties.'

[28] Sandeep Pandey, Sourashis Roy, Christopher Olston, 'Shuffling a Stacked Deck: The Case for Partially Randomized Ranking of Search Engine Results.' School of Computer Science Carnegie Mellon University Pittsburgh, 2005.

[29] Won-Seok Hwang, Soo-Min Chae, Sang-Wook Kim, 'Yet Another Paper Ranking Algorithm Advocating Recent Publications.' 2010.

[30] XAMPP, https://en.wikipedia.org/wiki/XAMPP.

[31] Αλέξανδρος Νίκας, 'Βελτιστοποίηση Ιστοσελίδων για Μηχανές Αναζήτησης.' Νοέμβριος 2011.

[32] Ηλίας Κανέλλος, 'Αναζήτηση σε επιστημονικές βάσεις δεδομένων με βάση την ιστορική εξέλιξη των δεδομένων.' Ιούλιος 2012.

[33] Γιώργος Φραγκιαδουλάκης, 'Υλοποίηση Μηχανής Αναζήτησης βασισμένης στο PageRank με χρήση του Hadoop.' Οκτώβριος 2013.