Ένα εργαλείο σε Java για την εξόρυξη πληροφοριών από ακαδημαϊκές εκδόσεις

Δημήτρης Ρουσίδης, Εμμανουήλ Γαρουφάλλου και Πάνος Μπαλατσούκας
drousid@gmail.com, garoufallou@gmail.com, pan-bal@hotmail.com

Περίληψη

Μια πληθώρα νέων δεδομένων και πληροφοριών προσθέτονται κάθε μέρα στον ακαδημαϊκό τομέα. Ο αριθμός των ακαδημαϊκών άρθρων αυξάνεται εκθετικά κάθε χρόνο καθιστώντας αυτή την τεράστια δεξαμενή γνώσης προβληματική ως προς την εξερεύνηση και επεξεργασία της. Προτείνεται ένα αυτοματοποιημένο εργαλείο εξόρυξης δεδομένων γραμμένο σε Java το οποίο θα εντοπίζει σημασιολογικές ομοιότητες μεταξύ ακαδημαϊκών συγγραμμάτων. Παρέχοντας στο εργαλείο έναν τίτλο άρθρου Α, θα μπορεί να αναγνωρίζει αποτελεσματικά άλλους τίτλους άρθρων τα οποία μοιράζονται κοινώς ένα ή παραπάνω κριτήρια, όπως συγγραφείς, αναφορές, λέξεις κλειδιά, θεματολογία, εκδότες, ημερομηνίες κα καθώς και μεθοδολογίες και θεματικές υποενότητες. Το εργαλείο θα έχει τη δυνατότητα ανακάλυψης κρυμμένων μοτίβων, κανόνων σχέσης (association rules), κατηγοριοποίησης (classification) και συσταδοποίησης-ομαδοποίησης (clustering) στις ακαδημαϊκές εκδόσεις καθώς και απεικόνιση όλων αυτών των πληροφοριών. Προκειμένου να διευκολυνθεί η ανακάλυψη των μεθοδολογιών, το προτεινόμενο εργαλείο θα είναι σε θέση να δημιουργεί μια βάση δεδομένων ορολογιών και υπο-ορολογιών κυρίως μέσω της ανάλυσης των ευρετηρίων ηλεκτρονικών βιβλίων (e-books). Επιπροσθέτως, αυτή η βάση δεδομένων θα είναι διαθέσιμη διαδικτυακά οπού ουσιαστικά θα δημιουργηθεί ένα αποθετήριο ορολογιών και υπο-ορολογιών. Ο κύριος σκοπός της δημιουργίας αυτού του εργαλείου είναι η διευκόλυνση της ακαδημαϊκής μελέτης και έρευνας και η αύξηση της ικανότητας άντλησης πληροφοριών και συσχετισμών από τα ακαδημαϊκά περιεχόμενα.

Λέξεις – Κλειδιά: Μεταδεδομένα, συστήματα υποστήριξης αποφάσεων, αναζήτηση πληροφοριών, ανάκτηση πληροφοριών, εξόρυξη γνώσης, εξόρυξη πληροφοριών.

1. Εισαγωγή

Πρόσφατες μελέτες έχουν δείξει ότι η αύξηση των επιστημονικών συγγραμμάτων είναι ραγδαία. Για παράδειγμα, οι Gargouri et al (2010) υπολόγισαν ότι ετησίως ο αριθμός των δημοσιευμένων άρθρων σε περιοδικά στα οποία έχει γίνει ομότιμη αναθεώρηση ήταν

περίπου 2,5 εκατ., ενώ ο Jinha (2010) εκτίμησε ότι ο συνολικός αριθμός των άρθρων που έχουν δημοσιευθεί σε περιοδικά μέχρι το 2009 ήταν πάνω από 50 εκατ.

Δεδομένου αυτού του πλούτου των συγγραμμάτων, η εύρεση των πιο σχετικών πληροφοριών γίνεται μια πολύ απαιτητική δουλειά όσο αφορά όχι μόνο στο χρόνο αλλά και τον κόπο από μέρους του ερευνητή. Παρόλο που οι σύγχρονες μηχανές αναζήτησης και μετα-αναζήτησης μπορούν να υποβοηθήσουν την αναζήτηση σχετικών πληροφοριών, υπάρχει ακόμα διαφωνία σχετικά με το κατά πόσο αυτά τα συστήματα μπορούν να ικανοποιήσουν τις ανάγκες και τις απαιτήσεις των χρηστών κατά τη διάρκεια της διαδικασίας αναζήτησης πληροφοριών. Για παράδειγμα, αρκετοί ερευνητές έχουν αναφέρει ότι οι σύγχρονες μηχανές αναζήτησης δεν υποστηρίζουν επαρκώς τις ανάγκες αυτών που αποζητούν σχετικότητα, κρίση και λήψη αποφάσεων στο Διαδίκτυο (Balatsoukas et al., 2009). Επιπλέον, οι γνωστικοί περιορισμοί του ανθρώπινου μυαλού δεν καθιστούν πάντα δυνατή την αξιολόγηση όλων των σχετικών πληροφοριών είτε λόγω της αναποτελεσματικότητας των επιλογών μας όσο αφορά στις στρατηγικές λογικής ή λόγω της τάσης μας να αισθανόμαστε ικανοποιημένοι με απλά αρκετές πληροφορίες (Prabha et al., 2007).

Ο σκοπός της παρούσας εργασίας είναι η αντιμετώπιση των προβλημάτων αυτών, προτείνοντας ένα νέο εργαλείο υποστήριξης αποφάσεων για την ανάκτηση και την ανάλυση ακαδημαϊκών συγγραμμάτων. Το προτεινόμενο εργαλείο κάνει χρήση της απεικόνισης μεταδεδομένων, της εξόρυξης κειμένου και της οπτικοποίησης πληροφοριών ως μέσα για την επαύξηση των ικανοτήτων λήψης αποφάσεων από τους ακαδημαϊκούς και τους φοιτητές κατά την αναζήτηση σχετικών επιστημονικών πληροφοριών. Η σύγκριση αυτή βασίζεται σε ένα σύνολο καθορισμένων από το χρήστη ποιοτικών κριτηρίων. Παρά το γεγονός ότι παρόμοιες προσπάθειες έχουν τεκμηριωθεί από ερευνητές στον τομέα της εξόρυξης αναφορών ανάκτησης πληροφοριών (π.χ. Kostoff, et al, 2001, Thelwall και Sud, 2011), αλλά και ομαδοποίησης άρθρων επιστημονικών περιοδικών (Delen και Crossland, 2008), η παρούσα μελέτη πηγαίνει ένα βήμα παραπέρα, εισάγοντας μια εκτενή λίστα επιστημονικών κριτηρίων που χρησιμοποιούνται για τη μέτρηση της ομοιότητας μεταξύ των άρθρων. Αυτά καταγράφονται σε μορφή σχεσιακού σχήματος μεταδεδομένων με σχετικές τιμές που ορίζονται για κάθε στοιχείο μεταδεδομένων. Μερικά μοναδικά μετα-δεδομένα που χρησιμοποιούνται για τη μέτρηση της ομοιότητας είναι εκείνα που σχετίζονται με τα εγγενή - ποιοτικά χαρακτηριστικά των επιστημονικών δημοσιεύσεων, όπως οι μέθοδοι που χρησιμοποιούνται για τη συλλογή και ανάλυση δεδομένων ή τον εντοπισμό των επιμέρους θεμάτων.

1.1 Βιβλιογραφική Ανασκόπηση

Σύμφωνα με την έκθεση του JISC (2012), υπάρχουν σήμερα πάνω από 144.000 πλήρους απασχόλησης επαγγελματίες στον ακαδημαϊκό χώρο (διδασκαλία και έρευνα) που εργάζονται στην τριτοβάθμια εκπαίδευση στο Ηνωμένο Βασίλειο. Μια εσωτερική έρευνα στην εταιρεία Elsevier, που προέρχεται από την ανάλυση του παγκόσμιου μοναδικού μετρητή χρηστών για την Science Direct, δείχνει ότι το συνολικό, παγκόσμιο αναγνωστικό κοινό περιοδικών μπορεί να είναι μεταξύ 10-15 εκατ. (Ware και Mabe, 2009). Οι εκτιμήσεις

της παγκόσμιας ερευνητικής κοινότητας καταρτίζονται από την UNESCO. Στην έκθεση της για το 2005 εκτιμήθηκε ότι υπάρχει παγκοσμίως μια βάση 5,5 εκατομμυρίων ερευνητών.

Η έκθεση RIN του 2008 από το Cambridge Economic Policy Associate όπως αναφέρεται στο άρθρο των Ware και Mabe (2009) υπολόγισε ότι το συνολικό κόστος της διεξαγωγής και κοινοποίησης της έρευνας που δημοσιεύεται σε περιοδικά είναι 175 δισ. λίρες, το οποίο αναλύεται σε 116 δισ. λίρες για τα έξοδα της ίδιας της έρευνας, 25 δισ. λίρες για τη δημοσίευση, διανομή και πρόσβαση στα άρθρα και 34 δισ. λίρες για την ανάγνωση τους. Ο Lamothe το 2012 διεξήγαγε ποσοτική ανάλυση εξετάζοντας τους παράγοντες που επηρεάζουν τη χρήση ηλεκτρονικών περιοδικών στη Βιβλιοθήκη JN Desmarais του Πανεπιστήμιο Laurentian (Καναδάς) και καλύπτει μια 11-ετή περίοδο από το 2000 έως το 2010. Τα αποτελέσματα της μελέτης έδειξαν μια εκθετική αύξηση του αριθμού των ηλεκτρονικών περιοδικών καθώς και λήψεων άρθρων.

Υπάρχει ένας σημαντικός αριθμός οφελών από την διαδικασία εξόρυξης κειμένου σύμφωνα με την έκθεση του JISC (2012): i) Εξοικονόμηση χρόνου: να κάνει κάποιος μια εργασία, όπως η ανασκόπηση βιβλιογραφίας, σε λιγότερο χρόνο από ότι κανονικά θα απαιτούνταν, ii) Βελτίωση της ποιότητας και της αξιοπιστίας των συμπερασμάτων: αύξηση κάλυψης της ύλης, iii) Αύξηση της παραγωγής: για παράδειγμα, περισσότερες ερευνητικές εργασίες, iv) Διαδικασία καινοτομίας: διευκόλυνση μιας εργασίας που διαφορετικά θα ήταν αδύνατη να πραγματοποιηθεί., v) Αποτελεσματικότητα, vi) Απελευθέρωση "κρυφών" πληροφοριών και ανάπτυξη νέων γνώσεων, vii) Διερεύνηση νέων οριζόντων, viii) Βελτίωση της έρευνας και της βάσης τεκμηρίωσης, ix) Βελτίωση της διαδικασίας έρευνας και της ποιότητας,

Οι συγγραφείς της ίδιας έκθεσης υλοποίησαν περιπτωσιολογικές μελέτες για να υποστηρίξουν την υπόθεσή τους, ότι η έρευνα με τη χρήση εργαλείων εξόρυξης κείμενου βελτιώνει τη διαδικασία παρέχοντας τεράστια εξοικονόμηση πόρων για τα πανεπιστήμια. Μία από τις περιπτωσιολογικές μελέτες απέδειξε ότι η αυτοματοποιημένη περίληψη μπορεί να υποστηρίξει την ανασκόπηση της βιβλιογραφίας. Σύμφωνα με τα ευρήματά τους:

Ο χρόνος που λαμβάνεται για να διαβάσει κάποιος ένα ακαδημαϊκό άρθρο συνοψίζοντας το περιεχόμενο του = 31 λεπτά
Χρονικό διάστημα για να διαβαστεί μια αυτοματοποιημένη περίληψη = 5 λεπτά,
Χρόνος που εξοικονομείται μέσω αυτοματοποιημένων περιλήψεων = 26 λεπτά
Έστω ότι ο μέσος ακαδημαϊκός μισθός £ 48.000 για 1.650 ώρες εργασίας ετησίως, τότε:
Εξοικονόμηση κόστους ανά περίληψη = £ 12,61

Σύμφωνα με τους Tenopir, King, Edwards και Wu (2009) ο μέσος ακαδημαϊκός επιστημών διαβάζει περίπου 204 μοναδικά άρθρα ανά έτος. Αν υποθέσουμε ότι η ίδια συμπεριφορά ανάγνωσης υφίσταται σε όλους τους κλάδους, η αυτοματοποιημένη περίληψη μέσω της εξόρυξης κειμένου θα μπορούσε να οδηγήσει σε εξοικονόμηση κόστους ανά ακαδημαϊκό ανά έτος ισοδύναμο με £ 2.572. Με πάνω από 144.000 ακαδημαϊκό προσωπικό στην ανώτατη εκπαίδευση της Βρετανίας, αυτό θα σημάνει ενδεχόμενη εξοικονόμηση έρευνας με απόδοση πάνω από £ 370 εκατομμύρια ετησίως.

1.2 Εργαλεία εξόρυξης κειμένου

Υπάρχει ένας αρκετά μεγάλος αριθμός εργαλείων λογισμικού εξόρυξης κειμένου. Η πλειοψηφία αυτών στοχεύει στην Αγγλική φιλολογία και τις Βιοϊατρικές επιστήμες. Το Sciverse, μια πρωτοποριακή πλατφόρμα για αξιόπιστο περιεχόμενο και εφαρμογές που επιταχύνουν την επιστημονική ανακάλυψη, διαθέτει μια συλλογή από 131 εργαλεία εξόρυξης κειμένου. Η στατιστική αξιολόγηση παρήγαγε τα ακόλουθα αποτελέσματα:

Τεχνική	**Αριθμός εργαλείων**
Αναγνώριση όρων	23
Στατιστική	8
Αναγνώριση στοιχείων (Πίνακες – Γραφήματα -Αποσπάσματα)	6
Κοινωνικά (Διαμοίραση – Επισήμανση – Μεταφορά - Αναφορά)	11
Αναζητήσεις	51
Σχετικά με συγγραφείς	10
Εξόρυξη κειμένου (Σχέσεις-Σχετικότητα-Συστάδες-Ταξινόμηση-)	14
Διαμορφώσεις (Κείμενο - Γραφήματα)	20

Πίνακας 1: *Αριθμός εργαλείων ανά τεχνική (κάποια εργαλεία έχουν παραπάνω από μια τεχνικές)*

Επιστημονικό Πεδίο	**Αριθμός εργαλείων**
Βιολογία - Ανθρωπολογικές επιστήμες	28
Ακαδημαϊκός τομέας	80
Υγεία (Ιατρικές επιστήμες)	16
Χημεία	3
Κοινωνικά Δίκτυα	3
Άλλο (πληροφορική, γεωργία, μετεωρολογία, πατέντες και εταιρείες)	5

Πίνακας 2: *Αριθμός πεδίων ανά επιστημονικό πεδίο (υπάρχουν εργαλεία με άνω του ενός Επ. Πεδ.)*

Από τα 131 εργαλεία λογισμικού, μόνο τα 22 (ποσοστό 16,8%) έχουν χρησιμοποιήσει είτε ανάλυση ή τεχνικές, όπως κανόνες σχέσης, κατηγοριοποίηση και ομαδοποίηση. Από αυτά τα 22 εργαλεία, τα 12 ανήκουν στο ακαδημαϊκό τομέα, τα 5 εκ οποίων επικεντρώνονται μόνο στους συντάκτες των άρθρων. Από τα εναπομείναντα 7 εργαλεία μόνο τα 2 προσεγγίζουν αμυδρά τη λογική του προτεινόμενου εργαλείου.

2. Σκοπός και στόχοι

Ο σκοπός της παρούσας εργασίας είναι να παρουσιάσει τα κύρια συστατικά ενός νέου εργαλείου λήψεως αποφάσεων για τη διεξαγωγή βιβλιογραφικής ανασκόπησης με βάση τα μεταδεδομένα. Ειδικότερα, ο σκοπός του εργαλείου είναι να κάνει χρήση της εξόρυξης κειμένου ως μέσο για τον εντοπισμό σημασιολογικών ομοιοτήτων μεταξύ των ακαδημαϊκών συγγραμμάτων (σε αυτό το στάδιο της έρευνας το εργαλείο επικεντρώνεται στην αξιολόγηση ανοικτής πρόσβασης άρθρων σε περιοδικά). Το εργαλείο μπορεί να συγκρίνει δύο ή περισσότερα άρθρα και να εμφανίζει ομοιότητες μεταξύ τους. Βασισμένο σε μεταδεδομένα για το άρθρο, τους συγγραφείς, τις λέξεις κλειδιά, τις αναφορές, την περίληψη και το κυρίως σώμα του άρθρου και ενός άρθρου «πηγή» (άρθρο το οποίο γνωρίζει ο ακαδημαϊκός ότι είναι υψηλής σημασίας για το αντικείμενό του), το εργαλείο υπολογίζει το ποσοστό ομοιότητας μεταξύ του άρθρου «πηγή» και οποιουδήποτε άλλου άρθρου Το εργαλείο μπορεί να

αποκαλύψει τις υποκατηγορίες και τις μεθόδους που χρησιμοποιούνται για τη συλλογή και ανάλυση δεδομένων για κάθε άρθρο. Επίσης, το εργαλείο είναι σε θέση να εφαρμόσει διάφορους σημαντικούς αλγορίθμους εξόρυξης δεδομένων για την ανακάλυψη κρυμμένων μοτίβων και πληροφοριών. Ειδικότερα, οι ειδικοί στόχοι του παρόντος εγγράφου είναι να συνοψίσει τα κύρια τεχνικά χαρακτηριστικά του εργαλείου και να εξηγήσει το ρόλο των μεταδεδομένων στη διαδικασία της λήψης αποφάσεων με βάση τα εργαλεία για βιβλιογραφικές ανασκοπήσεις.

Αυτό το έγγραφο είναι δομημένο ως εξής: Η επόμενη ενότητα παρουσιάζει μια περιγραφή των κύριων χαρακτηριστικών του εργαλείου. Η χρήση και ο ρόλος των μεταδεδομένων για την ανάπτυξη αυτού του εργαλείου παρουσιάζονται στην επόμενη ενότητα. Τέλος, το έγγραφο καταλήγει με ένα περίγραμμα των επόμενων βημάτων που εμπλέκονται στο σχεδιασμό και την αξιολόγηση του εργαλείου.

3. Μεθοδολογία - Περιγραφή των βασικών τεχνικών χαρακτηριστικών

Το εργαλείο αυτό βασίζεται στη Java. Αρχικά αναπτύχθηκε ως μια εφαρμογή υπολογιστών γραφείου. Ωστόσο, το εργαλείο θα μπορούσε να ενσωματωθεί σε έναν ιστοχώρο ή να χρησιμοποιείται ως πρόσθετο σε ήδη υπάρχοντα συστήματα ανάκτησης πληροφοριών (όπως ακαδημαϊκά αποθετήρια και ακαδημαϊκές ψηφιακές βιβλιοθήκες) και μηχανές αναζήτησης (πχ Google Scholar). Επίσης, όταν το λογισμικό έχει ολοκληρωθεί, θα δοκιμαστεί σε διαφορετικές πλατφόρμες, όπως τα Microsoft Windows, Linux και Mac OS X για να διασφαλιστεί ότι το πρόγραμμα λειτουργεί σωστά.

3.1 Χαρακτηριστικά και Μηχανισμός Εργαλείου

Αυτή η ενότητα θα παρέχει μια βήμα-προς-βήμα ανάλυση στην κύρια ροή εργασίας του εργαλείου. Αυτή περιλαμβάνει: μεταφόρτωση και μετατροπή των αρχείων, δημιουργία των πινάκων, επιλογή των κριτηρίων ομοιότητας και την εξόρυξη δεδομένων.

3.1.1 Μεταφόρτωση και μετατροπή των αρχείων

Αρχικά ένα αρχείο (πχ ένα άρθρο περιοδικού) θα πρέπει να τροφοδοτείται στο εργαλείο. Τα αρχεία που δέχεται το εργαλείο είναι όλα τα αρχεία κειμένου (.doc, .docx, .odt, .pdf, .rtf, .txt) με τη συντριπτική πλειοψηφία να έχουν pdf επέκταση. Το εργαλείο είναι προγραμματισμένο να επεξεργάζεται αρχεία με επέκταση xml αλλά τα δεδομένα εισόδου του χρήστη δεν χρειάζεται να έχουν αυτή την επέκταση. Το εργαλείο με αυτοματοποιημένη διαδικασία και χωρίς να έχει γνώση ο χρήστης, θα μετατρέψει αυτόματα όλα τα αρχεία που του μεταφορτώνονται σε μορφή xml. Η επιλογή της μορφής xml έγινε κυρίως γιατί έχει τα εξής πλεονεκτήματα: i) Η αναζήτηση των δεδομένων είναι εύκολη και αποτελεσματική. Οι μηχανές αναζήτησης μπορούν να αναλύσουν απλά την περιγραφή που φέρουν οι ετικέτες (tags) και να μην ανακατεύουν τα δεδομένα. Οι ετικέτες παρέχουν στις μηχανές αναζήτησης τη νοημοσύνη που τους λείπει, ii) Πολύπλοκες σχέσεις όπως τα δέντρα (trees) και η κληρονομικότητα (inheritance) μπορούν να κοινοποιηθούν, iii) Ο κώδικας είναι πολύ πιο

ευανάγνωστος ακόμα και για κάποιον ο οποίος έρχεται σε επαφή με το xml περιβάλλον, χωρίς προηγούμενη γνώση.

3.1.2 Δημιουργία πινάκων

Το επόμενο βήμα είναι η δημιουργία των πινάκων της βάσης δεδομένων για κάθε μεταφορτωμένο αρχείο. Ο αριθμός και το είδος των πινάκων βασίζεται σε μια σειρά προκαθορισμένων κριτηρίων. Τα κριτήρια αποτελούνται από ένα σύνολο μεταδεδομένων, με ειδικό λεξιλόγιο (κατά περίπτωση), που παρέχουν μια λεπτομερή περιγραφή των περιεχομένων του αρχείου. Ένα δείγμα αυτού του τύπου μεταδεδομένων παρουσιάζεται στον Πίνακα 3. Όπως και με τη χρήση των μεταδεδομένων, το εργαλείο είναι σε θέση να κατακερματίσει την περίληψη ή το κύριο σώμα του άρθρου ή και τα δύο, να αναλύσει και να δημιουργήσει νέους πίνακες με όλες τις λέξεις και τον αριθμό εμφάνισής τους στο άρθρο. Αυτό είναι ιδιαίτερα χρήσιμο σε περιπτώσεις όπου λέξεις-κλειδιά δεν είναι διαθέσιμα ως ξεχωριστό πεδίο στο άρθρο.

<table>
<tr>
<td>Μεταδεδομένα Άρθρου
Τίτλος άρθρου
Επιστημονικό πεδίο άρθρου
Εκδότης
Μορφή-επέκταση άρθρου
Τύπος άρθρου
Ανοικτή Πρόσβαση (Ναι / Όχι)
Έτος Έκδοσης
Ξεπερασμένο (Απόσταση από το τρέχον έτος)
Online (Ναι / Όχι)</td>
<td>Μεταδεδομένα Συγγραφέα
Αριθμός Συγγραφέων (Ακέραιος)
Όνομα-τα Συγγραφέα-ων (Χαρακτήρες)
Κατάταξη Συγγραφέα-ων (h-index, εάν υπάρχει)
Εργοδότης Συγγραφέα-ων (Χαρακτήρες, όνομα σχέσης)</td>
</tr>
<tr>
<td>Μεταδεδομένα Λέξεων-Κλειδιά
Λέξεις-Κλειδιά υπάρχουν στο άρθρο (Ναι / Όχι)
Χειροκίνητη εισαγωγή Λέξεων-Κλειδιά (Ναι / Όχι)
Αυτόματη αναζήτηση Λέξεων-Κλειδιά (Ναι / Όχι)
Αριθμός Λέξεων-Κλειδιά (Ακέραιος)</td>
<td rowspan="3">Μεταδεδομένα αναφορών
Αριθμός αναφορών (Ακέραιος)
Μετασχηματισμός (Ναι / Όχι)
Συγγραφέας αναφοράς (Χαρακτήρες)
Πηγή αναφοράς (Χαρακτήρες)
Έτος Έκδοσης (Ακέραιος)
Τίτλος αναφοράς (Χαρακτήρες)
Τόμος (Ναι / Όχι, Ακέραιος)
Αριθμός εγγράφου (Ναι / Όχι, Ακέραιος)
Εκδότες (Ναι / Όχι, Χαρακτήρες)
URL (Ναι / Όχι, Χαρακτήρες)
Ημερομηνία πρόσβασης (Ναι / Όχι, Ημερομηνία)
Τύπος (Εφημερίδα / Πρακτικά / Μονογραφία / Διαδίκτυο / Άλλα)</td>
</tr>
<tr>
<td>Μεταδεδομένα Περίληψης
Διαθέσιμη Περίληψη (Ναι / Όχι)
Ανάλυση για Λέξεις-Κλειδιά (Ναι / Όχι)
Κατακερματισμός Κειμένου (Ναι / Όχι)</td>
</tr>
<tr>
<td>Μεταδεδομένα Κυρίου Σώματος
Διαθέσιμο (Ναι / Όχι)
Ανάλυση για Λέξεις-Κλειδιά (Ναι / Όχι)
Κατακερματισμός Κειμένου (Ναι / Όχι)</td>
</tr>
</table>

Πίνακας 3: *Λίστα σχεσιακών μεταδεδομένων*

3.1.3 Επιλογή και ανάθεση των κριτηρίων ομοιότητας

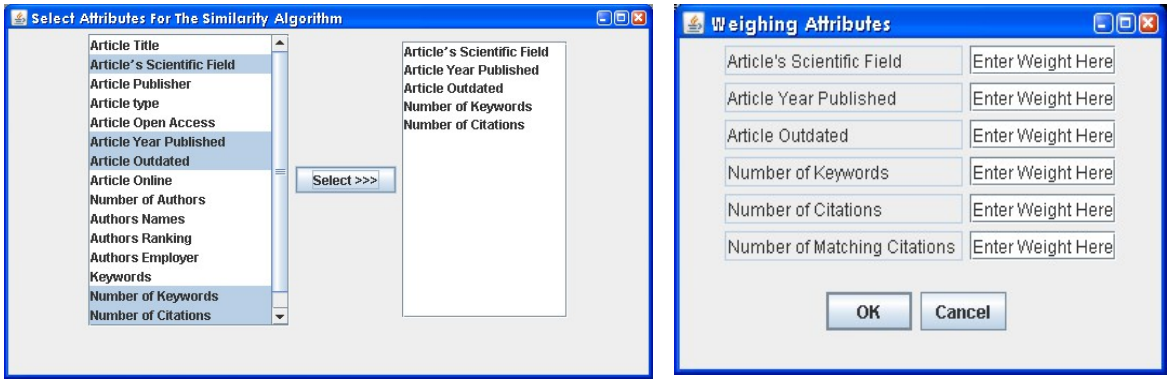
Δεν είναι όλα τα μεταδεδομένα της ίδιας βαρύτητας και σημασίας. Υπάρχουν στοιχεία που είναι πολύ πιο σημαντικά και πρέπει να δοθεί ιδιαίτερη προτεραιότητα και βαρύτητα κατά την εφαρμογή του αλγόριθμου ομοιότητας. Ως εκ τούτου διαφορετική βαρύτητα, θα ανατεθεί σε επιλεγμένα στοιχεία μεταδεδομένων. Η απόφαση αυτή ελήφθη διότι οι μελετητές και οι φοιτητές έχουν την τάση να παρουσιάζουν μια δυναμική συμπεριφορά κατά την αξιολόγηση της καταλληλότητας των επιστημονικών δεδομένων (Saracevic, 2007). Αυτή η

δυναμική συμπεριφορά, η οποία χαρακτηρίζεται από αλλαγές στον τρόπο που οι χρήστες τείνουν να εφαρμόσουν τα μεταδεδομένα για να αναζητήσουν και να αξιολογήσουν σχετικές πληροφορίες, είναι κοινή σε περιπτώσεις όπου αυτοί που αναζητούν πληροφορίες βιώνουν γνωστικές μεταβολές και αλλοιώσεις της ανώμαλης κατάστασης της γνώσης (Belkin, 1980).

Προκειμένου να αντιμετωπιστεί αυτή η δυναμική συμπεριφορά, το εργαλείο παρέχει στους χρήστες τη δυνατότητα να επιλέξουν τα πιο σημαντικά στοιχεία μεταδεδομένων και να εκχωρήσουν τη βαρύτητα σε αυτές τις ιδιότητες (Εικόνα 1). Η διαδικασία της ανάθεσης βαρύτητας στα επιλεγμένα στοιχεία μεταδεδομένων μπορεί να ολοκληρωθεί είτε χειροκίνητα ή αυτόματα. Στην περίπτωση της χειροκίνητης ανάθεσης βαρύτητας ζητείται από τον χρήστη να αξιολογήσει τη σημασία κάθε επιλεγμένου στοιχείου μεταδεδομένων χρησιμοποιώντας ένα δεκαδικό αριθμό, με ένα ψηφίο, μεταξύ 0 και 1. Επί του παρόντος, για την αυτόματη εκχώρηση της βαρύτητας, οι τιμές υπολογίζονται από το εργαλείο από την εξίσωση 1. Ωστόσο, η έρευνα βρίσκεται σε εξέλιξη, προκειμένου να πειραματιστεί με διαφορετικούς τύπους βαρύτητας. Αυτά είναι ενσωματωμένα σε ένα πρότυπο μέτρο ομοιότητας συνημίτονου, το προϊόν του οποίου χρησιμοποιείται για τον προσδιορισμό ομάδων από άρθρα που είναι παρόμοια με τις ανάγκες των χρηστών και μεταφράζει τις εγγραφές των πινάκων της βάσης δεδομένων σε υλικό που είναι κατάλληλο για την ομαδοποίηση (clustering) μέσω της εξόρυξης δεδομένων. Λόγω του μικρού μήκους του παρόντος εγγράφου μια πιο λεπτομερής περιγραφή του μέτρου ομοιότητας και των αλγόριθμων εξόρυξης δεδομένων θα παρουσιαστεί σε μετέπειτα δημοσιεύσεις.

$$W_i = \log \frac{N}{X_{ij}} \qquad \textit{(Εξίσωση 1)}$$

όπου: $\mathbf{W_i}$ = η βαρύτητα, $\mathbf{N}$ = ο αριθμός των άρθρων στη βάση δεδομένων και $\mathbf{X_{ij}}$ = ο αριθμός όλων των άρθρων (j) που μοιράζονται μια κοινή τιμή μεταδεδομένου με το άρθρο «πηγή» (i). Αυτή η τιμή θα μπορούσε να είναι μια ημερομηνία έκδοσης ή ένα σετ από λέξεις-κλειδιά.

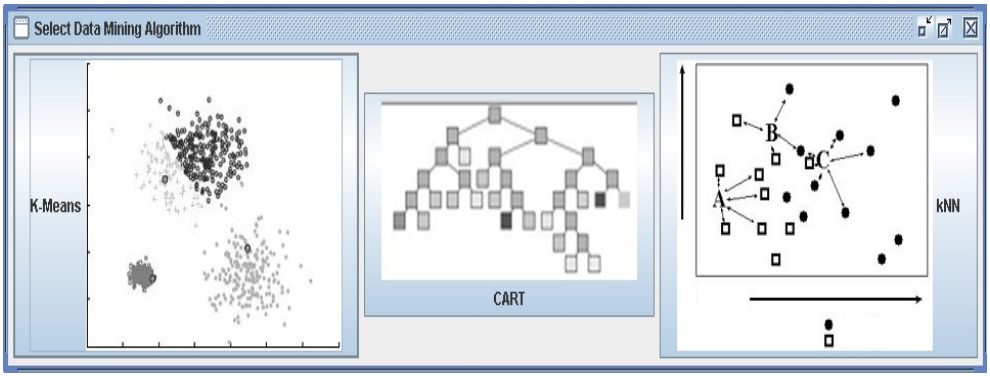


Εικόνα 1: *Επιλέγοντας και προσθέτοντας βαρύτητα στα στοιχεία.*

3.1.4 Επιλογή του αλγορίθμου εξόρυξης δεδομένων

Μετά τη δημιουργία των πινάκων και των πεδίων τους τουλάχιστον δύο αρχεία πρέπει να επεξεργαστούν από το εργαλείο έτσι ώστε επαρκή στοιχεία να είναι διαθέσιμα για την εφαρμογή των αλγορίθμων εξόρυξης δεδομένων. Ταξινόμηση (classification) και ανάλυση ομαδοποίησης (clustering analysis) θα πρέπει να εφαρμοστούν και οι πιο κατάλληλοι

αλγόριθμοι εξόρυξης δεδομένων να επιλεγούν. Το λογισμικό πρέπει να παρέχει επαρκή ποικιλία των τεχνικών εξόρυξης δεδομένων και αλγορίθμων για να επιλέξει ο χρήστης, οπότε μια πλειάδα αλγορίθμων θα μελετηθούν όπως: C4.5, K-Means, SVM: Support Vector Machines, EM, PageRank, AdaBoost, k-Nearest Neighbors, Naïve Bayes and CART: Classification and Regression Trees και οι πιο κατάλληλοι θα επιλεχθούν. Μια τυπική απεικόνιση της επιλογής εξόρυξης δεδομένων φαίνεται στην Εικόνα 2. Χρησιμοποιώντας τη λειτουργία της εξόρυξης δεδομένων, ο χρήστης θα πρέπει να είναι σε θέση να απεικονίσει το σύνθετο επιστημονικό περιβάλλον (όσον αφορά στην παραγωγή επιστημονικών άρθρων) και την κατασκευή μοντέλων που παρουσιάζουν διάφοροι τύποι τάσεων, όπως η πρόοδος των εργασιών του συγκεκριμένου επιστήμονα, μιας ομάδας επιστημόνων ή ενός συγκεκριμένου ακαδημαϊκού ιδρύματος στο πέρασμα του χρόνου, καθώς και τον προσδιορισμό των περιοχών αναδυόμενου ερευνητικού ενδιαφέροντος.



Εικόνα 2: *Επιλογή του data mining αλγορίθμου που θα εφαρμόσει το εργαλείο*

3.2 Ο ρόλος των μεταδεδομένων στη διαδικασία λήψης αποφάσεων για την κατασκευή εργαλείων για την ακαδημαϊκή ανασκόπηση

Το κομμάτι αυτό συνοψίζει το ρόλο των μεταδεδομένων. Ειδικότερα, το σχήμα της εικόνας 3 παρουσιάζει τη ροή εργασίας μεταδεδομένων που χρησιμοποιείται για να υποστηρίξει την ανάπτυξη και την εφαρμογή των μεταδεδομένων για τις ανάγκες αυτού του έργου.

3.2.1 Η διαδικασία της επιλογής, επικύρωσης και ορισμού

Η επιλογή των στοιχείων των μεταδεδομένων βασίστηκε σε βιβλιογραφική ανασκόπηση στο χώρο έρευνας για τη συμπεριφορά στην αναζήτηση πληροφοριών. Ειδικότερα, η ανασκόπηση επικεντρώθηκε σε έγγραφα που καταδεικνύουν τα κριτήρια που οι άνθρωποι τείνουν να εφαρμόσουν κατά την αναζήτηση και την αξιολόγηση της ακαδημαϊκών εκδόσεων. Παραδείγματα των κριτηρίων αυτών περιλαμβάνουν: Επικαιρότητα (δηλαδή λέξεις-κλειδιά, θεματικές επικεφαλίδες, περιλήψεις), Ποιότητα (δηλαδή στοιχεία για το συγγραφέα, την υπαγωγή του συγγραφέα ή τη φήμη του συγγραφέα), Σαφήνεια (δηλαδή παρουσία πινάκων, στοιχείων, ακατέργαστων συνόλων δεδομένων), Εγκυρότητα (πχ καταλληλότητα των μεθόδων που χρησιμοποιούνται για τη συλλογή δεδομένων), ή Σχέση με άλλες πηγές (πχ χρήση των αναφορών, τύπος των αναφορών, τύπος των συγγραφέων που παρατίθενται). Τα κριτήρια αυτά είχαν μετατραπεί σε αντίστοιχα στοιχεία μεταδεδομένων. Κάθε στοιχείο μεταδεδομένων συνοδεύτηκε από την αντίστοιχη τιμή (Πίνακας 3). Όπως και με την επισκόπηση της βιβλιογραφίας, ο πίνακας παρουσιάστηκε για επικύρωση στο σύνολο

των βιβλιοθηκονόμων και επιστημόνων του προσωπικού του ΤΕΙ Θεσσαλονίκης. Έχοντας προσδιορίζει το βασικό σύνολο των στοιχείων των μεταδεδομένων, το επόμενο βήμα ήταν ο καθορισμός τους που βασίζεται σε XML σχήμα μεταδεδομένων. Αυτό το σχήμα χρησιμοποιήθηκε ως πρότυπο για τη δημιουργία των πινάκων της βάσης δεδομένων.

3.2.2 Δημιουργία των πινάκων μεταδεδομένων

Κάθε πίνακας περιέχει πληροφορίες σχετικά για ένα συγκεκριμένο στοιχείο μεταδεδομένων. Τα στοιχεία μεταδεδομένων είναι σχεσιακά επειδή οι πίνακες μπορούν να συνδεθούν και έτσι κληρονομούν δεδομένα που περιέχονται μεταξύ των διαφόρων στοιχείων μεταδεδομένων. Η σχεσιακή ένωση γίνεται με τη χρήση των χαρακτηριστικών που είναι κοινά στους διάφορους πίνακες μεταδεδομένων. Για παράδειγμα, η απλή ιδιότητα ID_ΑΡΧΕΙΟΥ θα μπορούσε να χρησιμοποιηθεί για τη σύνδεση του πίνακα ΑΡΧΕΙΟ με τον πίνακα ΛΕΞΕΙΣ. Με αυτόν τον τρόπο το τελικό αποτέλεσμα των σχεσιακών πινάκων μεταδεδομένων θα μπορούσε να θεωρηθεί ως μια σημασιολογική δομή του ιστού ενός ακαδημαϊκού σύμπαντος (στην περίπτωση του ερευνητικού αυτού έργου αυτό το σύμπαν περιορίζεται στα σύνολα δεδομένων των τίτλων των περιοδικών που τροφοδοτούνται στο πρωτότυπο εργαλείο).

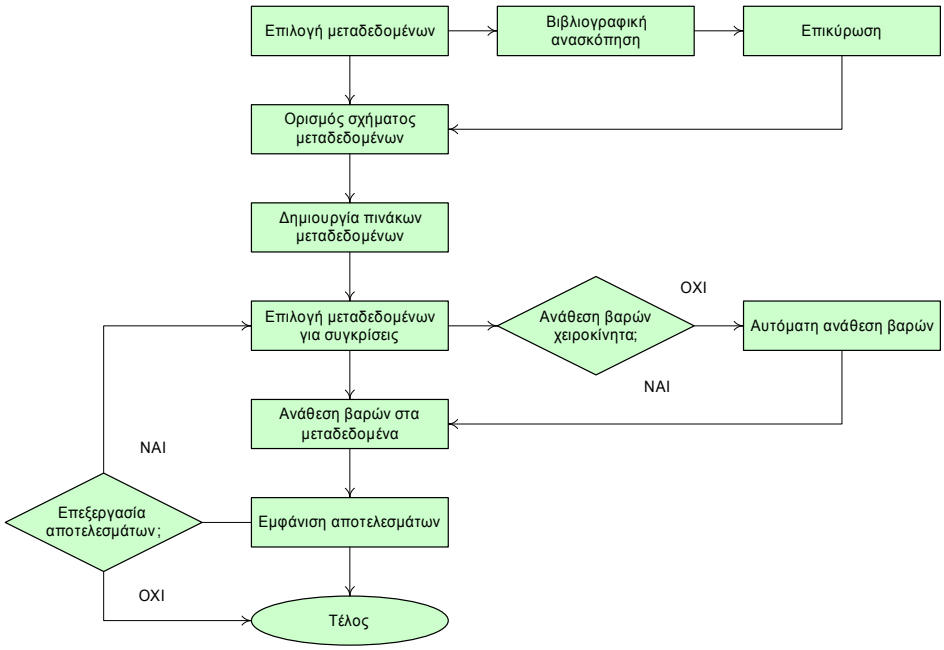
3.2.3 Απόδοση των συγκρίσεων ομοιότητας

Ο καθορισμός των κριτηρίων για τη διενέργεια συγκρίσεων ομοιότητας μεταξύ περιοδικών άρθρων αποτελεί βασική συνιστώσα της αλληλεπίδρασης των χρηστών με το πραγματικό σύνολο μεταδεδομένων. Συγκεκριμένα, ο χρήστης έχει τη δυνατότητα να επιλέξει τα στοιχεία μεταδεδομένων που πρέπει να ληφθούν υπόψη από τον αλγόριθμο εξόρυξης δεδομένων. Επίσης, ο χρήστης μπορεί να ορίσει χειροκίνητα την βαρύτητα για κάθε στοιχείο μεταδεδομένων. Η χειροκίνητη ρύθμιση του βάρους είναι απλή, δίνοντας την ευκαιρία στον χρήστη να ορίσει οποιαδήποτε τιμή μεταξύ 0 και 1 (0 = χαμηλή σημασία ενώ 1 = μεγάλη σημασία). Επίσης, ο χρήστης μπορεί να επιλέξει ανάμεσα σε χειροκίνητη, αυτόματη ή ημι-αυτόματη εκχώρηση της βαρύτητας. Στην περίπτωση της ημι-αυτόματης επιλογής, ο χρήστης μπορεί να αποφασίσει ποια μεταδεδομένα πρέπει να χρησιμοποιηθούν, απλά χρησιμοποιώντας μπάρες κύλισης και δίνοντας στα μεταδεδομένα που τον ενδιαφέρουν ονομαστικούς χαρακτηρισμούς όπως αδιάφορο, σχεδόν αδιάφορο, λίγο σημαντικό, σημαντικό, υποχρεωτικό. Σε αυτούς τους ονομαστικούς χαρακτηρισμούς είναι προκαθορισμένη η βαρύτητα από το εργαλείο. Ως εκ τούτου, σε αντίθεση με τα συνήθη εργαλεία ανάκτησης πληροφοριών που αντιμετωπίζουν ισότιμα όλα τα μεταδεδομένα, το παρόν εργαλείο επιτρέπει στους χρήστες να αντιστοιχίσουν βαθμούς σημασίας για ένα επιλεγμένο σύνολο μεταδεδομένων.

3.2.4 Χειραγώγηση των αποτελεσμάτων / απόδοση των νέων συγκρίσεων ομοιότητας

Δύο τύποι αποτελεσμάτων θα πρέπει να παρουσιάζονται στην ενότητα αποτελεσμάτων: 1) ένας ιεραρχημένος κατάλογος όλων των άρθρων που είναι παρόμοια με αυτά που χρησιμοποιούνται για την αρχικοποίηση της σύγκρισης, και 2) μια οπτική αναπαράσταση του αποτελέσματος (για παράδειγμα, με βάση την ομαδοποίηση K-means) (Εικόνα 2). Ο χρήστης έχει τη δυνατότητα να χειραγωγήσει τα αποτελέσματα. Αυτό επιτυγχάνεται είτε με την έναρξη μιας νέας σύγκρισης από την αρχή, ή με τη μεταβολή του υφισταμένου συνόλου αποτελεσμάτων. Στην τελευταία περίπτωση, αυτό μπορεί να επιτευχθεί με την τροποποίηση

των μεταδεδομένων. Για παράδειγμα, η βαρύτητα των μεταδεδομένων μπορεί να τροποποιηθεί ως μέσο για την εκ νέου λειτουργία του αλγορίθμου εξόρυξης δεδομένων και ελέγχου για το νέο σύνολο των αποτελεσμάτων. Επίσης, πρόσθετα στοιχεία από τα μεταδεδομένα μπορούν να επιλεγούν για την ένταξη στις συγκρίσεις ομοιότητας. Επιπλέον, ο χρήστης έχει τη δυνατότητα να χρησιμοποιήσει μεταδεδομένα για να χτίσει μοντέλα που αντιπροσωπεύουν τάσεις στην επιστημονική δημοσίευση, πχ ο χρήστης μπορεί να επιλέξει να παρουσιάσει αλλαγές στο ιστορικό δημοσιεύσεων ενός ή περισσότερων επιστημόνων. Αυτές οι αλλαγές μπορεί να είναι είτε αλλαγές στο βασικό θέμα, σε υπο-θέματα, μεθόδους που χρησιμοποιούνται ή ακόμη και σε τίτλους περιοδικών που χρησιμοποιούνται για να διασπείρουν τη δουλειά του. Η εργασία είναι σε εξέλιξη για τη θέσπιση απαιτήσεων των χρηστών σχετικά με το είδος των μοντέλων που θα μπορούσαν να χρησιμοποιηθούν για διάφορα προφίλ χρηστών (πχ φοιτητές, μεταπτυχιακοί φοιτητές, είτε ακαδημαϊκοί).



Εικόνα 3: *Ροή εργασίας μεταδεδομένων*

3.3 Προσδιορισμός μεθοδολογιών ακαδημαϊκών άρθρων

Ο κύριος στόχος του δεύτερου μέρους του εργαλείου είναι να μπορεί να προσδιορίσει τις μεθοδολογίες που αναφέρονται στα κύρια μέρη των άρθρων. Για να επιτευχθεί αυτή η λειτουργία πρέπει να δημιουργηθεί μια τεράστια βάση δεδομένων γλωσσάριων και υπο-γλωσσάριων. Οι ορολογίες θα μπορούν να είναι ενσωματωμένες στο εργαλείο ή να μεταφορτώνονται από το χρήστη ή να δημιουργούνται εκ του μηδενός μέσω του εργαλείου αναλύοντας ευρετήρια από ηλεκτρονικά βιβλία (e-books).

3.3.1 Δημιουργία γλωσσάριων και υπο-γλωσσάριων

Η κατηγοριοποίηση των γλωσσάριων και των υπό-γλωσσάριων θα βασίζεται πάνω στη δενδροειδή δομή. Τα κυρίως γλωσσάρια θα προκύψουν από τα κύρια επιστημονικά πεδία όπως η Πληροφορική, Ιατρική, Φυσικές Επιστήμες, κλπ. Προκειμένου να επιτευχθούν τα καλύτερα δυνατά αποτελέσματα πρέπει να υλοποιηθεί μια τεράστια βάση δεδομένων

γλωσσάριων. Συνεπώς, προτείνεται μια αυτοματοποιημένη διαδικασία που θα είναι σε θέση να αναλύσει τα ευρετήρια των e-books και να δημιουργήσει τα γλωσσάρια βασιζόμενη στη σελίδα και το κεφάλαιο στα οποία βρίσκεται η κάθε λέξη που συναντάται στα ευρετήρια των ηλεκτρονικών βιβλίων.

Κάθε ηλεκτρονικό βιβλίο θα μετατραπεί σε ένα αρχείο XML και χρησιμοποιώντας όλα τα μεταδεδομένα και τα χαρακτηριστικά των αρχείων XML, όπως ετικέτες (labels), κεφαλίδες (headers), κ.λπ. οι λέξεις από το ευρετήριο του e-book μπορούν να ανατεθούν σε συγκεκριμένα υπο-γλωσσάρια τα οποία θα προκύψουν ανάλογα με την παράγραφο, την ενότητα και το κεφάλαιο στα οποία ανήκει η κάθε λέξη.

Κάθε λέξη θα φέρει ετικέτες από τα υπο-γλωσσάρια στα οποία ανήκει, το καθένα με διαφορετική βαρύτητα και όσο πιο απομακρυσμένο είναι από την κορυφή του δέντρου Ιεραρχίας, άρα και όσο πιο εξειδικευμένο είναι, τόσο μεγαλύτερο θα είναι η βαρύτητα της ετικέτας.

Το εργαλείο εφαρμογής αναλύοντας το κάθε ακαδημαϊκό άρθρο θα είναι σε θέση να προσδιορίσει τα υπο-γλωσσάρια στα οποία η κάθε λέξη ανήκει. Ανιχνεύοντας και συνδυάζοντας τα υπο-γλωσσάρια το εργαλείο θα έχει μια ισχυρή ένδειξη της μεθοδολογίας που ακολουθείται από το συγγραφέα.

3.4 Δοκιμή και Αξιολόγηση

Μια τεράστια αποθήκη ακαδημαϊκών άρθρων, περιοδικών, πρακτικών και e-books δημιουργείται και θα αριθμεί μερικές χιλιάδες ακαδημαϊκά συγγράμματα. Η κύρια πηγή για τη συσσώρευση όλου αυτού του ηλεκτρονικά δημοσιευμένου υλικού προέρχεται από ανοικτής πρόσβασης ηλεκτρονικές βιβλιοθήκες, καταλόγους, βάσεις δεδομένων διατριβών, πανεπιστημιακές βιβλιοθήκες και ακαδημαϊκές προσωπικές συλλογές. Η αξιολόγηση του εργαλείου θα γίνει μέσω των τεστ χρηστικότητας και μελετών χρηστών από φοιτητές, πανεπιστημιακούς, αλλά και με εκτιμήσεις ειδικών.

4. Συμπεράσματα – Τρέχουσα εργασία

Προτείνεται ένα εργαλείο που διερευνά τη σκοπιμότητα της εφαρμογής αλγορίθμων εξόρυξης δεδομένων, προκειμένου να ανακαλυφθεί η ομοιότητα μεταξύ των ακαδημαϊκών εγγράφων που βασίζεται σε καινοτόμα ποιοτικά κριτήρια . Μέχρι σήμερα, η πλειονότητα των εργαλείων ασχολήθηκε με την εξέταση της ομοιότητας μόνο από την άποψη της τοπικότητας και άλλων αντικειμενικών κριτηρίων (όπως το όνομα του συγγραφέα, υπαγωγή, κλπ). Ωστόσο, το προτεινόμενο εργαλείο πηγαίνει ένα βήμα παραπέρα βασιζόμενο σε άλλα πιο εγγενή και ουσιαστικά χαρακτηριστικά της έρευνας που σχετίζονται με τη χρησιμοποιούμενη μέθοδο ή την ανάλυση των δεδομένων.

Η έρευνα είναι σε εξέλιξη για τη βελτίωση και περαιτέρω ανάπτυξη του εργαλείου. Ειδικότερα, όσον αφορά το σχεδιασμό του συστήματος, η ομάδα έργου επικεντρώνεται σε

περαιτέρω πειραματισμούς με διάφορους αλγόριθμους ομοιότητας και βαρύτητα μεταδεδομένων. Όσον αφορά την διεπαφή χρήστη-εργαλείου, δοκιμές χρηστικότητας βρίσκονται σε εξέλιξη, προκειμένου να προσδιοριστούν τα κύρια προβλήματα.

Βιβλιογραφία

Balatsoukas, P., Morris, A., and O'Brien, A. (2009). An evaluation framework of user interaction with metadata. Journal of Information Science, vol. 35, no.3, pp. 321-339.

Belkin, N. J. (1980). Anomalous styate of knowledge as a basis for information retrieval. *The Canadian journal of Information Science*, vol. 5, pp.133-143

Delen D., Crossland M. D., (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications*, pp 1707-1720.

Gargouri, Y, Hajjem, C, Lariviere, V, Gingras, Y, Brody, T, Carr, L and Harnad, S. (2010). Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research. *PLOS ONE*, 5, (10). http://www.plosone.org/article/info:doi/10.1371/journal.pone.0013636 (Πρόσβαση: 13/07/2012)

Jinha, A., 2010. Article 50 million: an estimate of the number of scholarly articles in existence. *LearnedPublishing, 23* (3), 258-263

JISC, 2012. *Value and benefits of text mining.* [online] Διαθέσιμο στο: <http://www.jisc.ac.uk/publications/reports/2012/value-and-benefits-of-text-mining.aspx#a01> (Πρόσβαση: 10/07/2012).

Kostoff, R., del Río, J., Humenik, J., García, E., and Ramírez, A., 2001. Citation Mining: Integrating Text Mining and Bibliometrics for Research User Profiling. *Journal of the American Society for Information Science and Technology 52*(13), pp.1148-1156.

Lamothe, A., 2012. Factors Influencing Usage of an Electronic Journal Collection at a Medium-Size University: An Eleven-Year Study. *Partnership: the Canadian Journal of Library and Information Practice and Research*, Volume 7, issue 1. ISSN: 1911-9593 (digital)

Prabha, C., Connaway, L., Olszewski, L. and Jenkins, L (2007). What is enough? Satisficing information needs. Journal of Documentation, vol. 63, no.1, pp.74-89.

Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. Journal of the American Society for Information Science and Technology, vol. 58, no.13, pp.2126-2144.

Thelwall, M. & Sud, P. (2011). A comparison of methods for collecting web citation data for academic organisation. *Journal of the American Society for Information Science and Technology*, 62(8), 1488–1497

Ware, M. and Mabe, M., 2009. *The stm report: An overview of scientific and scholarly journal publishing.* [online] Διαθέσιμο στο: <http://www.stm-assoc.org/2009_10_13_MWC_STM_Report.pdf