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

Πληροφοριακό Σύστημα Ενημέρωσης Δικτύου Βιβλιογραφικών Αναφορών από τον Ιστό με τεχνικές εξαγωγής πληροφορίας, τεχνολογίας λογισμικού και ταιριάσματος όμοιων εγγραφών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΓΕΩΡΓΙΟΥ Α. ΠΑΠΑΔΑΚΗ

Επιβλέπων : Τιμολέων Σελλής
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2007

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

Πληροφοριακό Σύστημα Ενημέρωσης Δικτύου Βιβλιογραφικών Αναφορών από τον Ιστό με τεχνικές εξαγωγής πληροφορίας, τεχνολογίας λογισμικού και ταιριάσματος όμοιων εγγραφών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΓΕΩΡΓΙΟΥ Α. ΠΑΠΑΔΑΚΗ

Επιβλέπων : Τιμολέων Σελλής,
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 8η Οκτωβρίου 2007.

.....	.....	.....
Τιμολέων Σελλής	Ιωάννης Βασιλείου	Νεκτάριος Κοζύρης
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2007

.....
ΓΕΩΡΓΙΟΣ Α. ΠΑΠΑΔΑΚΗΣ

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

Copyright © Παπαδάκης Γεώργιος, 2007
Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

Η αξιολόγηση του ερευνητικού τους έργου και ο προσδιορισμός της απήχησης που αυτό έχει απασχολούσε ανέκαθεν τους επιστήμονες. Στα μέσα της δεκαετίας του 1950 προτάθηκε για το σκοπό αυτό η μέθοδος των βιβλιογραφικών αναφορών (citations), η οποία στις μέρες μας είναι καθολικά αποδεκτή ως η πιο αξιόπιστη. Απαιτείται ωστόσο ιδιαίτερη προσπάθεια και κόπος για να καταφέρει κανείς να συγκεντρώσει τα citations για όλες τις δημοσιευμένες εργασίες του, ακόμα και στην σημερινή εποχή του Διαδικτύου. Χρειάζεται άλλωστε να συνδυάσει πληροφορίες από πλήθος ετερογενών πηγών. Είναι επομένως επιτακτική η ανάγκη για αυτοματοποίηση της διαδικασίας αυτής. Από τις ιδιαίτερα αξιόλογες προσπάθειες που έχουν γίνει προς αυτή την κατεύθυνση, καμία δεν έχει καταφέρει να λύσει επιτυχώς το σύνολο των προβλημάτων που πρέπει να αντιμετωπίσει μια προσπάθεια αυτοματοποίησης. Σε αυτά συγκαταλέγονται η ελεύθερη πρόσβαση και επεξεργασία (parsing) των πρωτογενών πηγών πληροφοριών (εκδοτικοί οίκοι κλπ), το ταίριασμα των διαφορετικών βιβλιογραφικών αναφορών που αναφέρονται στην ίδια δημοσίευση (citation matching) και ο εντοπισμός των διαφορετικών επιστημόνων που συμμετέχουν στη συγγραφή ενός συνόλου δημοσιεύσεων (name disambiguation). Για την ακρίβεια, το name disambiguation επιμερίζεται στον εντοπισμό εκείνων των ονομάτων που, παρ' όλο που ταυτίζονται, αντιστοιχούν στην πραγματικότητα σε διαφορετικούς επιστήμονες (mixed citation problem) και των ονομάτων που, παρ' όλο που διαφέρουν, αντιστοιχούν στην πραγματικότητα στον ίδιο επιστήμονα (split citation problem). Αντικείμενο αυτής της διπλωματικής είναι η ανάπτυξη, με βάση τις αρχές της τεχνολογίας λογισμικού, ενός συστήματος ανάλυσης βιβλιογραφικών αναφορών που αντιμετωπίζει το σύνολο των παραπάνω προβλημάτων (στην περίπτωσή μας βέβαια το πρώτο πρόβλημα ανάγεται στην εξαγωγή πληροφορίας από ακαδημαϊκές μηχανές αναζήτησης). Σε αυτό το πλαίσιο αναπτύχθηκαν πρωτότυποι αλγόριθμοι για την επίλυση τόσο του citation matching όσο και του name disambiguation. Οι αλγόριθμοι αυτοί βασίζονται στις τεχνικές ταιριάσματος όμοιων εγγραφών και διαφοροποιούνται από τους προτεινόμενους στη βιβλιογραφία, καθώς δεν προορίζονται για εφαρμογή σε ένα περιορισμένο σύνολο δεδομένων. Αντίθετα, στόχος είναι να χρησιμοποιηθούν σε μια εφαρμογή πραγματικού χρόνου, ώστε να επιτυγχάνουν σε αποδεκτό χρόνο υψηλή απόδοση σε οποιαδήποτε δεδομένα. Αν και είναι εξαιρετικά δύσκολο να υπολογιστεί η αποτελεσματικότητα και η αξιοπιστία μιας τέτοιας εφαρμογής, τα πρώτα αποτελέσματα είναι ικανοποιητικά, αφήνοντας παράλληλα αρκετά περιθώρια βελτίωσης.

Λέξεις κλειδιά: εξαγωγή πληροφορίας, citation matching, name disambiguation, mixed citation problem, split citation problem, string distance metrics

Abstract

The evaluation of their research work and its effect has always been one of scholars' great concerns. In the middle 50s a new evaluation method based on citations was proposed. Nowadays this method is widely accepted as the most reliable one. However, gathering a scholar's citations constitutes a particularly laborious task, even in the current Internet era, as one needs to correctly combine information from miscellaneous sources. There exists therefore an urgent need for automating this process. Numerous remarkable efforts have been made to cover this need, none of which has managed though to solve all related problems. Among these problems are the necessary free access to and parsing of primary information sources (e.g. publishers), citation matching (i.e. indentifying citations that actually refer to the same paper) and name disambiguation. Name disambiguation refers to the task of identifying the unique authors that contribute to a set of papers and is further separated to the following problems: mixed citation problem, caused by the fact that different scholars may have identical names, and split citation problem, induced by the various names under which a unique scholar appears. This thesis aims to develop an information system according to software engineering principles that copes with all of the aforementioned problems (in our case however the first one is reduced to extracting information from academic search engines). In this context we developed algorithms that deal with citation matching as well as name disambiguation based on record linkage techniques and string distance metrics. These algorithms differ from those proposed so far in literature in that they are not appropriate just for a limited dataset. They are instead made appropriate for a real time application that should process any data successfully and in a reasonable time. Although estimating the accuracy and reliability of such an application is a fairly difficult task, the first results are encouraging enough. Moreover, we plan to improve the algorithms and enrich the application with some necessary new features in the future.

Λέξεις κλειδιά: information extraction, citation matching, name disambiguation, mixed citation problem, split citation problem, string distance metrics

Ευχαριστίες

Κίνητρο για να αναλάβω τη συγκεκριμένη διπλωματική ήταν η δυνατότητα που μου παρείχε να εφαρμόσω τις σημαντικότερες γνώσεις που αποκόμισα όλα αυτά τα χρόνια από τη ΣΗΜΜΥ (προγραμματισμός, βάσεις δεδομένων, τεχνολογίες διαδικτύου) και μάλιστα στα πλαίσια της τεχνητής νοημοσύνης. Μετά από αρκετούς μήνες ενασχόλησης, μπορώ με βεβαιότητα να πω ότι τα οφέλη που αποκόμισα από αυτή την εργασία είναι πέρα από τις αρχικές μου προσδοκίες. Άλλωστε ήρθα σε μια πρώτη επαφή με δύο μέχρι πρότινος άγνωστα σε μένα αντικείμενα, την μηχανική μάθηση και κυρίως την έρευνα. Για το λόγο αυτό αλλά και για την άριστη συνεργασία που είχαμε όλους αυτούς τους μήνες θα ήθελα να ευχαριστήσω τον ερευνητή του Ε.Κ.Ε.Φ.Ε "Δημόκριτος" Γεώργιο Παλίουρα.

Θα ήθελα επίσης να ευχαριστήσω θερμά τον καθηγητή Τίμο Σελλή για τη δυνατότητα που μου έδωσε να αναλάβω αυτή τη διπλωματική, για τη βοήθεια που μου παρείχε όποτε τη χρειάστηκα αλλά και γενικότερα για την υποδειγματική του στάση ως καθηγητής.

Θα ήθελα ακόμα να ευχαριστήσω τους φίλους και συμφοιτητές, Αφροδίτη Κυρλιγκίτση και Θάνο Παπαοικονόμου, για την αμέριστη συμπαράστασή και βοήθειά τους και σε αυτήν την προσπάθειά μου. Τέλος, από τις ευχαριστίες δεν θα μπορούσα να παραλείψω την οικογένειά μου που με στηρίζει όλα αυτά τα χρόνια.

Πίνακας Περιεχομένων

1 Εισαγωγή 1
1.1 Γενικά 1
1.2 Αντικείμενο της διπλωματικής εργασίας 3
1.3 Διάρθρωση της εργασίας 5
2 Υπάρχουσες Προσεγγίσεις 7
2.1 Δημοσιοποίηση των ερευνητικών εργασιών στον Παγκόσμιο Ιστό 7
2.1.1 Scopus 8
2.1.2 Web of Science 9
2.1.3 Google Scholar 10
2.1.4 Λοιπές αξιόλογες ακαδημαϊκές μηχανές αναζήτησης 12
2.1.5 Συμπεράσματα για τις ακαδημαϊκές μηχανές αναζήτησης 13
2.2 Αυτοματοποίηση της συλλογής και ταύτισης βιβλιογραφικών εγγραφών 14
2.2.1 Βασικές έννοιες μηχανικής μάθησης 14
2.2.2 Mixed & Split Citation Problem 15
2.2.3 Citation Matching Problem 20
2.2.4 Μετρικές Μορφολογικής Απόστασης Συμβολοσειρών 22
2.2.5 Εξαγωγή πληροφορίας (Information Extraction) & Wrapper Maintenance . 24
3 Ανάκτηση δεδομένων από το Web , Επίλυση των προβλημάτων CMP, MCP & SCP 27
3.1 Εύρεση δημοσιεύσεων και πιθανών συνωνύμων για το δοσμένο όνομα επιστήμονα 28
3.2 Ανάκτηση των δημοσιεύσεων που αντιστοιχούν σε κάθε επιβεβαιωμένο συνώνυμο 30
3.3 Επίλυση του Citation Matching προβλήματος 31
3.4 Επεξεργασία των δημοσιεύσεων από το χρήστη και καταχώρησή τους στη βάση δεδομένων 35
3.5 Επίλυση των προβλημάτων mixed citation και split citation 37
3.6 Επεξεργασία των διαφορετικών επιστημόνων από το χρήστη και καταχώρησή τους στη βάση δεδομένων 43
3.7 Αναζήτηση citations για μια δημοσίευση 45
3.8 Αξιολόγηση Αλγορίθμων 46
4 Ανάπτυξη Πληροφοριακού Συστήματος 49
4.1 Επιλογή Μοντέλου Κύκλου Ζωής 50
4.2 Προσδιορισμός των απαιτήσεων 50
4.2.1 Λειτουργικές απαιτήσεις 51
4.2.2 Μη λειτουργικές απαιτήσεις 51
4.3 Αρχιτεκτονική Σχεδίαση (Architectural Design) 52
4.3.1 Γραφικό Περιβάλλον (GUI Package) 53
4.3.2 Επικοινωνία με Βάση Δεδομένων (DBMS Package) 55
4.3.3 Διαχείριση πληροφοριών (InfoManagement Package) 57
4.3.4 Wrapper του Google Scholar (GSWrapper Package) 59
4.3.4 Μέθοδοι Μηχανικής Μάθησης (MachineLearning Package) 64
5 Επίλογος 65
5.1 Σύνοψη και Συμπεράσματα 65
5.2 Μελλοντικές Επεκτάσεις 66
Βιβλιογραφία 69

1

Εισαγωγή

1.1 Γενικά

Η αξιολόγηση του ερευνητικού έργου ενός επιστήμονα συνιστά από τη φύση της μια εξαιρετικά δύσκολη εργασία. Και αυτό γιατί η ποσοτικοποίηση μιας αφηρημένης, μη μετρίσιμης έννοιας, όπως η συνεισφορά ενός ερευνητή στην επιστήμη, αποτελεί μια αμφιλεγόμενη διαδικασία. Κάθε φιλόδοξη μέθοδος αξιολόγησης πρέπει, επομένως, να χαρακτηρίζεται τουλάχιστον από αντικειμενικότητα, ενώ, όπως θα δούμε και στη συνέχεια, εξίσου επιθυμητή θεωρείται και η δυνατότητα αυτοματοποίησης της.

Πριν προχωρήσουμε, πρέπει να επισημάνουμε ότι θεωρούμε ότι η **επιστημονική εργασία** ενός ερευνητή συνίσταται στη δημοσιοποίηση των πονημάτων του με έναν από τους ακόλουθους τρόπους :

- Βιβλίο ή Κεφάλαιο σε Βιβλίο (*Book Chapter*)
- Δημοσίευση σε αναγνωρισμένο Επιστημονικό Περιοδικό (*Journal Paper*)
- Δημοσιοποίηση στα πλαίσια αναγνωρισμένου Επιστημονικού Συνεδρίου (*Conference Paper*)
- Διδακτορική Διατριβή (*Ph.D. Thesis*)
- Διπλωματική Εργασία (*M.Sc. Thesis*) ή Προπτυχιακή εργασία (*B.Sc. Thesis*), οι οποίες όμως δεν αφορούν πάντα πρωτότυπες εργασίες
- Τεχνική Αναφορά (*Technical Report*)
- Ανάρτηση σε έγκυρο ακαδημαϊκό ή επιστημονικό (*.edu*) Ιστοτόπο του World Wide Web
- Κατοχυρωμένη Ευρεσιτεχνία (*patent*)

Από τις μεθόδους αξιολόγησης που κατά καιρούς προτάθηκαν, επικράτησε η μέθοδος των βιβλιογραφικών αναφορών. Πρωτοπόρος στην καθιέρωση της μεθόδου αυτής ήταν ο **Eugene Garfield**, με την πρώτη σχετική εισήγησή του να χρονολογείται από το 1955 ("*Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas*"). Η

μέθοδος αυτή ορίζει ότι ([WI07]) η αξία του έργου ενός επιστήμονα είναι ανάλογη του συνολικού αριθμού των βιβλιογραφικών αναφορών που αφορούν στις δημοσιευμένες εργασίες του ή του μέσου αριθμού βιβλιογραφικών αναφορών για κάθε δημοσιοποιημένη εργασία του.

Οι **βιβλιογραφικές αναφορές** (*citations*) ([WI07]) αποτελούν μια μέθοδο αναγνώρισης της συνεισφοράς μιας επιστημονικής εργασίας σε κάποιο άλλο ερευνητικό έργο. Προς τούτο οφείλουν να περιέχουν επαρκή στοιχεία για τον ορθό εντοπισμό της συγκεκριμένης επιστημονικής πηγής. Γενικά, οι βιβλιογραφικές αναφορές παρατίθενται στο τέλος μιας δημοσίευσης, ενός κεφαλαίου ή βιβλίου και πρέπει να είναι συμβατές με κάποιους καθιερωμένους κανόνες. Στην *πιο διαδεδομένη τους μορφή*, δηλαδή όσον αφορά τους τρεις πρώτους από τους παραπάνω τρόπους δημοσιοποίησης, περιλαμβάνουν (με την σειρά που αναφέρονται):

- Όνομα συγγραφέα *ή* ονόματα συγγραφέων σε φθίνουσα σειρά βαθμού συμμετοχής στην εργασία
- Τίτλο Εργασίας *ή* Βιβλίου
- Τίτλο Περιοδικού (*journal*) *ή* Συνεδρίου (*conference*) όπου δημοσιοποιήθηκε η εργασία *ή* Όνομα Εκδοτικού Οίκου (*publisher*) που εξέδωσε το βιβλίο
- Τόμο (*volume*), Έκδοση (*issue*), Σελίδες (*pages*) Περιοδικού/Πρακτικών Συνεδρίου
- Έτος Δημοσιοποίησης

Αξίζει να σημειωθεί ότι οι βιβλιογραφικές αναφορές σχετίζονται άμεσα με άλλες, πιο σύνθετες μετρικές για την *άμεση* αξιολόγηση τόσο επιστημόνων, όπως ο **H-index** που προτάθηκε το 2005 από τον Jorge E. Hirsch ([WI07]) , όσο και επιστημονικών περιοδικών, όπως ο **Impact Factor** (IF) ([WI07]) που προτάθηκε από τον E. Garfield. Ωστόσο, αν και αποτελεί μια αρκετά διαδεδομένη μετρική, ο IF αμφισβητείται έντονα, ενώ ομοίως αμφιλεγόμενος είναι και ο H-index. Περισσότερες πληροφορίες για τις μετρικές αυτές αλλά και γενικότερα για δημοσιεύσεις σχετικές με το αντικείμενο μπορεί να βρει κανείς στην ιστοσελίδα της *Thomson ISI*, της εταιρείας που διαχειρίζεται το έργο του E. Garfield (http://scientific.thomson.com/free/essays).

Δεν λείπει λοιπόν και η κριτική όσον αφορά την αντικειμενικότητα της μεθόδου των βιβλιογραφικών αναφορών, η οποία εστιάζεται στα εξής ζητήματα:

- Οι επιστήμονες που εμπλέκονται στην εκπόνηση μιας εργασίας, δεν καταβάλλουν συνήθως ισάξια προσπάθεια. Αυτό άλλωστε αντικατοπτρίζεται και στη σειρά με την οποία αναγράφονται τα ονόματά τους όταν αυτή δημοσιευθεί. Ωστόσο, ο συνολικός αριθμός βιβλιογραφικών αναφορών ενός επιστήμονα δεν λαμβάνει υπόψη του αυτή τη διαβάθμιση.

- Υπάρχει ο κίνδυνος να στηριχθεί μια δημοσίευση σε κάποια άλλη εργασία χωρίς τελικά να την αναφέρει στη βιβλιογραφία της. Το φαινόμενο αυτό ονομάζεται διεθνώς ***plagiarism*** - *λογοκλοπή*.
- Ο αριθμός των βιβλιογραφικών αναφορών ενός επιστήμονα μπορεί να διογκωθεί από της αυτοαναφορές (***self-citations***), διαστρεβλώνοντας τον πραγματικό αριθμό.
- Μια βιβλιογραφική αναφορά δεν συνεπάγεται πάντα ότι το έργο ενός επιστήμονα συνέβαλε σε αυτό ενός άλλου. Αντίθετα, δεν είναι σπάνιο το φαινόμενο μια βιβλιογραφική αναφορά να παρατίθεται με μόνο στόχο να ενημερώσει τον αναγνώστη για παρεμφερείς δημοσιεύσεις ή ακόμα και για να αμφισβητηθούν τα αποτελέσματα της ([Broo86]).

Παρατηρούμε, επομένως, ότι υπάρχουν αρκετές ενστάσεις σχετικά με τη χρήση των βιβλιογραφικών αναφορών ως μέθοδο αξιολόγησης της ερευνητικής διαδικασίας. Δεδομένης όμως της μεγάλης δυσκολίας και περιπλοκότητας του συγκεκριμένου ζητήματος, είναι εύλογη και αποδεκτή η ύπαρξη ορισμένων μειονεκτημάτων σε μια τέτοια μέθοδο. Συνεπώς, υπό το πρίσμα της αντικειμενικότητας, η λύση των βιβλιογραφικών αναφορών φαντάζει αρκετά αξιόπιστη. Όσον αφορά το θέμα της αυτοματοποίησης της, σε αυτό είναι αφιερωμένη η επόμενη ενότητα.

1.2 Αντικείμενο της διπλωματικής εργασίας

Η καταμέτρηση των συνολικών βιβλιογραφικών αναφορών για ένα επιστήμονα έχει απλοποιηθεί σε μεγάλο βαθμό τα τελευταία χρόνια, κυρίως χάρη στον Παγκόσμιο Ιστό. Παραμένει ωστόσο μια χρονοβόρα και επιρρεπής στα λάθη διαδικασία, λόγω της ανάγκης συνδυασμού πληροφοριών από διάφορες πηγές του Web. Για το λόγο αυτό καθίσταται επιτακτική η αυτοματοποίησή της. Υπάρχουν όμως αρκετά προβλήματα που πρέπει να επιλυθούν από μία προσπάθεια αυτοματοποίησης, μεταξύ των οποίων σημαντικότερα είναι τα εξής:

- <u>Πρόσβαση στις πρωτογενείς πηγές δημοσιοποίησης επιστημονικών εργασιών,</u> δηλαδή σε πρακτικά συνεδρίων, άρθρα περιοδικών, βιβλία έγκυρων εκδοτικών οίκων κλπ. Η εργασία αυτή απαιτεί την υλοποίηση ενός συστήματος που θα μετατρέπει τις πληροφορίες που είναι διαθέσιμες από *ετερογενείς* πηγές σε μια *κοινή* μορφή. Εκτός αυτού όμως απαραίτητη είναι και η σύναψη συμφωνιών με τους εκάστοτε κατόχους πνευματικών δικαιωμάτων για ελεύθερη πρόσβαση στο περιεχόμενο των εκδόσεων τους. Επίσης, δεν πρέπει να παραληφθεί το γεγονός ότι αρκετές δημοσιεύσεις

υπάρχουν δυστυχώς μόνο σε έντυπη και όχι σε ηλεκτρονική μορφή, γεγονός που παρεμβάλει ένα ακόμα εμπόδιο, αυτό της ψηφιοποίησης.

- Ταίριασμα Βιβλιογραφικών Αναφορών (*Citation Matching* - ***CMP***). Αιτία του προβλήματος αυτού είναι η διαφορετική μορφή που έχουν οι βιβλιογραφικές αναφορές που αναφέρονται στην ίδια δημοσίευση. Κύριος παράγοντας αυτής της ασυμφωνίας είναι η έλλειψη πρότυπης μορφής, με την οποία να είναι συμβατές όλες οι βιβλιογραφικές αναφορές, ανεξαρτήτως επιστημονικού κλάδου.
- Πρόβλημα των Ομώνυμων Συγγραφέων (*Mixed Citation Problem* – ***MCP***). Ο (άκομψος στα ελληνικά) αυτός όρος χρησιμοποιείται για να περιγράψει την κατάσταση κατά την οποία διαφορετικοί επιστήμονες έχουν τα ίδια ακριβώς ονόματα. Σαν αποτέλεσμα, είναι δύσκολη η αντιστοίχηση μιας δημοσίευσης στο σωστό επιστήμονα κατά τον υπολογισμό των βιβλιογραφικών αναφορών τους .
- Πρόβλημα των Συνώνυμων Συγγραφέων (*Split Citation Problem* - ***SCP***). Το πρόβλημα αυτό δημιουργείται στις περιπτώσεις που το όνομα του ίδιου επιστήμονα εμφανίζεται με διαφορετικές μορφές σε διαφορετικές βιβλιογραφικές αναφορές, ακόμα και αν αυτές αναφέρονται στην ίδια δημοσίευση. Στα αίτια του φαινομένου αυτού συγκαταλέγονται διάφοροι λόγοι, όπως ορθογραφικά λάθη.

Προς το παρόν, δεν υπάρχουν κοινά αποδεκτές λύσεις για τα παραπάνω προβλήματα, παρόλο που η σχετική βιβλιογραφία είναι πλούσια. Επιπλέον, όσες προσπάθειες έχουν γίνει σχετικά με το συγκεκριμένο αντικείμενο είτε επικεντρώνονται σε ένα μόνο από τα προβλήματα αυτά είτε περιορίζονται σε καθαρά πειραματικούς σκοπούς. Περιορίζονται δηλαδή στην ανάπτυξη μεθόδων οι οποίες εφαρμόζονται σε επιλεγμένα και περιορισμένα σύνολα δεδομένων που δεν προορίζονται για χρήση σε εφαρμογές πραγματικού χρόνου. Περισσότερες λεπτομέρειες για τις αντίστοιχες δημοσιεύσεις και τις σημαντικότερες προσπάθειες επίλυσης τους παρατίθενται στο δεύτερο κεφάλαιο. Επισημαίνουμε επίσης ότι για ευκολία στο εξής θα αναφερόμαστε στα προβλήματα αυτά με τους αγγλικούς όρους μόνο.

Σκοπός της διπλωματικής αυτής εργασίας είναι να προτείνει και να εφαρμόσει μια ενιαία μέθοδο επίλυσης των παραπάνω προβλημάτων στα πλαίσια μιας εφαρμογής που θα αντλεί τα απαιτούμενα δεδομένα από το Web. Στην περίπτωσή μας δηλαδή το πρώτο πρόβλημα εκφυλίζεται στην επεξεργασία (*parsing*) των αποτελεσμάτων μιας ακαδημαϊκής μηχανής αναζήτησης, η οποία έχει *απευθείας* πρόσβαση στις σχετικές πηγές. Πιο συγκεκριμένα, η εργασία αυτή επικεντρώνεται στην ανάπτυξη ενός απλού πληροφοριακού συστήματος το οποίο θα παρέχει στο χρήστη τις εξής δυνατότητες:

- να αναζητήσει τις δημοσιεύσεις ενός επιστήμονα σε συνδυασμό με τα citations τους από πηγές στον Παγκόσμιο Ιστό και να καταχωρήσει τα αποτελέσματα σε μία βάση δεδομένων για μετέπειτα επεξεργασία. Η ορθότητα των πληροφοριών που

καταχωρούνται στη ΒΔ εξασφαλίζεται από τους αλγορίθμους μηχανικής μάθησης που αναπτύξαμε για την επίλυση των προβλημάτων MCP, SCP και CMP, σε συνδυασμό με την επιβεβαίωση των αποτελεσμάτων τους από το χρήστη.

- να εμφανίσει τα καταχωρημένα δεδομένα με τον επιθυμητό τρόπο (ταξινόμηση ανά αριθμό citations, ταξινόμηση ανά χρονολογία δημοσίευσης) και να εντοπίσει άγνωστες πληροφορίες που προκύπτουν από την επεξεργασία τους. Ειδικότερα προβλέπεται η εύρεση όλων των συν-συγγραφέων (**co-authors**) ενός επιστήμονα αλλά και η εύρεση όλων των συγγραφέων που έχουν αναφέρει ένα τουλάχιστον paper του δοσμένου συγγραφέα (**citing authors**).

Επιπλέον, έγινε προσπάθεια να συμβαδίσει η ανάπτυξη της εφαρμογής με τις προσταγές της σύγχρονης τεχνολογίας λογισμικού, ώστε μεταξύ άλλων να είναι εύκολη η συντήρηση της εφαρμογής αλλά και η μελλοντική επέκτασή της με νέες δυνατότητες. Χρήσιμες δυνατότητες για παράδειγμα θα ήταν αφενός η οπτικοποίηση του γράφου του κοινωνικού δικτύου – *social network* που δημιουργούν οι επιστήμονες μέσω των citations και αφετέρου η εξαγωγή των citations ενός επιστήμονα σε αρχείο απλού κειμένου ή/και σε xml αρχείο.

1.3 Διάρθρωση της εργασίας

Το υπόλοιπο της εργασίας οργανώνεται ως ακολούθως:

- Στο κεφάλαιο 2 παρουσιάζονται οι πιο αντιπροσωπευτικές προσπάθειες επίλυσης καθενός από τα προηγούμενα προβλήματα.
- Στο κεφάλαιο 3 γίνεται λεπτομερής αναφορά στα διαδοχικά στάδια τα οποία περιλαμβάνει μια αναζήτηση δεδομένων από το Web. Με άξονα τη διαδικασία αυτή παρουσιάζονται οι μέθοδοι μηχανικής μάθησης που αναπτύξαμε για την επίλυση των παραπάνω προβλημάτων καθώς και τα σημεία στα οποία εφαρμόζονται.
- Το κεφάλαιο 4 είναι αφιερωμένο στη σχεδίαση του συστήματος, δηλαδή τον διαχωρισμό του σε υποσυστήματα. Από αυτά αναπτύσσονται εκτενώς εκείνα που σχετίζονται με την επεξεργασία των δεδομένων από το Web, τη διαχείριση και παρουσίαση των καταχωρημένων δεδομένων καθώς και το γραφικό περιβάλλον.
- Στο κεφάλαιο 5 παρουσιάζονται τα συμπεράσματα της εργασίας και προτείνονται θέματα που αξίζει να διερευνηθούν στο μέλλον.

2

Υπάρχουσες Προσεγγίσεις

Στα πλαίσια του κεφαλαίου αυτού θα αναφερθούμε στις πιο αξιόλογες από τις λύσεις που έχουν προταθεί στη βιβλιογραφία για καθένα από τα προαναφερθέντα προβλήματα που αντιμετωπίζει μια προσπάθεια αυτόματης καταγραφής βιβλιογραφικών αναφορών. Τα ζητήματα αυτά τα ομαδοποιήσαμε σε δυο κατηγορίες, τόσο διαφορετικές μεταξύ τους που θα μπορούσαν να αναπτυχθούν σε ξεχωριστά κεφάλαια. Η πρώτη από αυτές αφορά τις ακαδημαϊκές μηχανές αναζήτησης, χάρη στις οποίες το πρόβλημα της πρόσβασης στις πρωτογενείς πηγές δημοσιοποίησης επιστημονικών εργασιών εκφυλίζεται στην περίπτωσή μας στην εξαγωγή πληροφορίας από τα αποτελέσματά τους. Η δεύτερη κατηγορία περιλαμβάνει τις μεθόδους μηχανικής μάθησης που έχουν αναπτυχθεί για την επίλυση των προβλημάτων citation matching, mixed citation και split citation. Για λόγους πληρότητας, όμως, επιλέξαμε να ενσωματώσουμε τις δυο αυτές κατηγορίες σε ένα ενιαίο κεφάλαιο, με την κάθε μια να καταλαμβάνει μια διαφορετική ενότητα.

2.1 Δημοσιοποίηση των ερευνητικών εργασιών στον Παγκόσμιο Ιστό

Από τα πρώτα χρόνια της άνθησης των δικτύων υπολογιστών και κυρίως της συνεχώς αυξανόμενης εξάπλωσης του Διαδικτύου εκδηλώθηκε έντονο ενδιαφέρον από εκπαιδευτικούς και μη οργανισμούς για την on-line διάθεση σε ψηφιακή μορφή του ερευνητικού έργου που παράγουν. Έτσι, μέχρι σήμερα έχουν υλοποιηθεί αναρίθμητες σχετικές προσπάθειες για τη δημιουργία των επονομαζόμενων **Ψηφιακών Βιβλιοθηκών** (*Digital Libraries*), σε τέτοιο βαθμό που και η απλή καταμέτρησή τους να είναι εξαιρετικά δυσχερής, καθώς ο αριθμός τους αυξάνεται συνεχώς. Μια εργασία που συνοψίζει τις πιο αξιόλογες πηγές και μηχανές αναζήτησης ακαδημαϊκών εργασιών είναι η [Zill07]. Πρέπει ωστόσο να σημειώσουμε το

γεγονός ότι ορισμένες από αυτές δεν παρέχουν ελεύθερη πρόσβαση στο περιεχόμενο τους παρά μόνον κατόπιν συνδρομής.

Αυτό που ενδιαφέρει, όμως, δεν είναι οι μεμονωμένες προσπάθειες, η πλειοψηφία των οποίων άλλωστε έχει περιορισμένη εμβέλεια και δυνατότητες (π.χ. απουσιάζει η δυνατότητα αναζήτησης βιβλιογραφικών αναφορών). Αντίθετα, σημαντικότερες είναι οι διαδικτυακές υπηρεσίες που προσφέρουν τη δυνατότητα για *ενιαία* πρόσβαση στο σύνολο αυτών των *ετερογενών* on-line πηγών σε συνδυασμό με τη δυνατότητα *ανάλυσης βιβλιογραφικών αναφορών*. Πράγματι, προς την κατεύθυνση αυτή έχουν αναπτυχθεί τα τελευταία χρόνια αξιόλογες *ακαδημαϊκές μηχανές αναζήτησης* και μάλιστα με καλές προοπτικές βελτίωσης. Οι ακαδημαϊκές μηχανές αναζήτησης υπερτερούν στον τομέα τους έναντι και των γενικών μηχανών αναζήτησης, στα εξής κυρίως θέματα ([Godi07]):

- *Αξιόπιστα αποτελέσματα*, καθώς οι πηγές από όπου αντλούνται οι πληροφορίες ελέγχονται για την εγκυρότητά τους. Με αυτόν τον τρόπο αντιμετωπίζεται επιτυχώς η διάχυτη καχυποψία που επικρατεί για την ποιότητα των πληροφοριών που είναι διαθέσιμες στο Διαδίκτυο.
- *Ακριβή αποτελέσματα*. Κάνοντας, δηλαδή, ο χρήστης μια αναζήτηση με έναν επιστημονικό όρο δεν θα λάβει ως αποτέλεσμα κείμενα όπου ο όρος αυτός χρησιμοποιείται με εντελώς διαφορετική έννοια (εκτός και αν χρησιμοποιείται ο ίδιος όρος και σε άλλη επιστημονική περιοχή). Χαρακτηριστικό παράδειγμα αποτελεί ο ιατρικός όρος Rapid Eye Movement που στη συντομευμένη του μορφή (REM) είναι όνομα μουσικού συγκροτήματος. Επίσης, στην ακρίβεια των αποτελεσμάτων συμβάλει και ο αποκλεισμός από τις αναζητήσεις των forum συζητήσεων.

Εκτενής αναφορά στις καλύτερες ακαδημαϊκές μηχανές αναζήτησης βασιζόμενη στα [BMR06], [Burr06], [Dess06] και [Fing06] γίνεται στις επόμενες υποενότητες.

2.1.1 Scopus - http://www.scopus.com/

Η υπηρεσία αυτή διατίθεται κατόπιν συνδρομής από το 2004. Δημιουργός της είναι η Elsevier Publishing Co, ολλανδικός εκδοτικός οίκος διεθνών επιστημονικών περιοδικών. Η βάση δεδομένων της αντλεί πληροφορίες κατεξοχήν από επιστημονικά περιοδικά (*journals*) σε ποσοστό 95%, ενώ πρόσθετες πηγές πληροφοριών αποτελούν οι ιστοσελίδες (*author homepages, university sites*) και οι πατέντες (patents). Η συνεισφορά των βιβλίων στο συνολικό αριθμό εγγραφών της ΒΔ (περίπου 30 εκατομμύρια) ανέρχεται μόλις στο 0.1%, ενώ εξίσου μικρή είναι και η αναλογία των εργασιών που δημοσιεύονται σε συνέδρια. Οι εγγραφές αυτές, που αυξάνονται κάθε χρόνο με γοργούς ρυθμούς, καλύπτουν τους εξής επιστημονικούς κλάδους (με φθίνουσα σειρά κάλυψης) :

- ιατρικές επιστήμες (*life and health sciences*)
- επιστήμες μηχανικών (*engineering*)
- βιολογία, γεωλογία & επιστήμες περιβάλλοντος (*biological, agricultural and environmental sciences*)
- λοιπές θετικές επιστήμες (*chemistry, physics, mathematics*)
- κοινωνικές επιστήμες, ψυχολογία και οικονομία (*social sciences, psychology and economics*)

Πρέπει να επισημάνουμε ότι η κάλυψη των δυο τελευταίων κατηγοριών είναι δυσανάλογα μικρή σε σχέση με τις τρεις πρώτες, ενώ μηδαμινή είναι η κάλυψη των κλάδων που συγκαταλέγονται στην κατηγορία "τέχνη και ανθρωπιστικές επιστήμες" (*art & humanities*), όπως η κλασσική φιλολογία (*classics*).

Όσον αφορά στη χρονολογική κάλυψη των εγγραφών της, οι μισές από αυτές αφορούν δημοσιεύσεις από το 1996 και μετά, ενώ οι υπόλοιπες αφορούν (θεωρητικά) δημοσιεύσεις από το 1900 (στην πράξη από τα μέσα του 1960).

Στα υπέρ της υπηρεσίας αυτής είναι το γεγονός ότι υποστηρίζει και δημοσιεύσεις σε γλώσσες εκτός των αγγλικών (εφόσον όμως το abstract είναι στα αγγλικά), ενώ από τα δυνατότερα σημεία της είναι η εύχρηστη διεπαφή χρήστη καθώς επίσης και οι πολυάριθμες επιλογές για βοήθεια. Έτσι ακόμα και αρχάριοι χρήστες μπορούν να τη χρησιμοποιήσουν αποδοτικά, ενώ παρέχει πολλές δυνατότητες παραμετροποίησης των αποτελεσμάτων αναζήτησης για προχωρημένους χρήστες, όπως χρήση boolean τελεστών και περιορισμούς όσον αφορά διάφορα πεδία (π.χ. χρονολογία, είδος εγγράφου και επιστημονικό πεδίο).

Τέλος, διαθέτει δυνατότητα αναζήτησης βιβλιογραφικών αναφορών, η οποία είναι ιδιαίτερα γρήγορη και με μεγάλες δυνατότητες ταξινόμησης των αποτελεσμάτων, Μοναδικό μελανό σημείο αποτελεί το γεγονός ότι τα αποτελέσματα της περιορίζονται χρονικά στην περίοδο από το 1996 και έπειτα.

Περισσότερες πληροφορίες παρέχονται στην ιστοσελίδα http://info.scopus.com.

2.1.2 Web of Science - http://www.ekt.gr/wos/index.html

Η υπηρεσία αυτή αποτελεί την κορωνίδα της εξέλιξης των προσπαθειών του E. Garfield και της εταιρείας Thomson ISI για 40 και πλέον χρόνια. Για το λόγο αυτό άλλωστε χαρακτηρίζεται από μεγάλη εξειδίκευση στην ανάλυση βιβλιογραφικών αναφορών, παρέχοντας πολλές σχετικές δυνατότητες, περισσότερες από κάθε άλλη αντίστοιχη υπηρεσία.

Στη σημερινή της μορφή, που διατίθεται μόνο κατόπιν συνδρομής, καλύπτει το σύνολο των επιστημονικών κλάδων. Το περιεχόμενό της το αντλεί από 5 εξειδικευμένες ΒΔ ή, αναλυτικότερα, από περίπου 9000 διεθνείς επιστημονικές επιθεωρήσεις αλλά και πρακτικά

περίπου 4500 επιστημονικών και τεχνικών συνεδρίων. Υποστηρίζει και άλλου είδους πηγές πληροφοριών, όπως πατέντες, αλλά δεν λαμβάνει υπόψη της καθόλου πηγές διαθέσιμες ελεύθερα στον Παγκόσμιο Ιστό, όπως προσωπικά sites επιστημόνων (*homepages*). Πρέπει, επιπλέον, να επισημάνουμε ότι καλύπτει πολύ λιγότερες μη αγγλόφωνες πηγές πληροφοριών σε σχέση με τους ανταγωνιστές της.

Σχετικά με την χρονολογική κάλυψη των εγγραφών της (συνολικά 40 εκατομμύρια), δεν παρέχονται επίσημες πληροφορίες αλλά έχει αποδειχτεί από πειράματα ότι υπερτερεί του Scopus για τα χρόνια πριν το 1996. Υστερεί όμως για την περίοδο από το 1996 και έπειτα. Μάλιστα, εμφανίζεται να επιστρέφει αποτελέσματα μέχρι και από το 1945 (ακόμα και από το 1900 για περίπου 200 περιοδικά).

Η υπηρεσία ανάλυσης βιβλιογραφικών αναφορών (το δυνατότερο σημείο του WoS) επιστρέφει αποτελέσματα ισάριθμα σε γενικές γραμμές με αυτά του Scopus. Εκτός αυτού στην πλειονότητα των επιστημονικών κλάδων παρατηρείται και μεγάλη ταύτιση των αποτελεσμάτων (της τάξεως του 80% με 90%). Ωστόσο, η υπηρεσία αυτή διαφοροποιείται από τους ανταγωνιστές της με τις προσφερόμενες επιλογές για *προχωρημένη ανάλυση βιβλιογραφικών αναφορών* (π.χ. για ολόκληρους οργανισμούς).

Τέλος, όσον αφορά στη διεπαφή χρήστη, υστερεί αισθητά στην ευχρηστία συγκρινόμενη με το Scopus. Προσφέρει ωστόσο παραπλήσιο πλήθος επιλογών παραμετροποίησης των αποτελεσμάτων της. Οι σύνθετες επιλογές επιβαρύνουν όμως σημαντικά την ταχύτητα απόκρισης της, η οποία σε γενικές γραμμές είναι μικρότερη από αυτή του Scopus.

2.1.3 Google Scholar - http://scholar.google.com/

Η διάθεση της υπηρεσίας αυτής στο ευρύ κοινό ξεκίνησε στις αρχές του 2005 από την Google. Αποτελεί δηλαδή μια καινούρια και αναπόφευκτα ανώριμη υπηρεσία, ιδιαίτερα αν αναλογιστεί κανείς πως δημιουργήθηκε εκ του μηδενός (σε αντίθεση π.χ. με το Scopus που στηρίχθηκε σε προϋπάρχουσες εξειδικευμένες μηχανές αναζήτησης της Elsevier όπως Geobase, Biobase, και Embase). Για το λόγο αυτό τα αποτελέσματά της σε ορισμένες περιπτώσεις δεν είναι ιδιαίτερα ποιοτικά, πράγμα για το οποίο δέχεται συχνά δριμεία κριτική (π.χ. [Jasc05], [Jasc06]).

Όσον αφορά στην εμβέλειά της, καλύπτει πρακτικά όλους τους ακαδημαϊκούς κλάδους καθώς επίσης και όλων των ειδών τις δημοσιεύσεις (με εξαίρεση τις πατέντες). Πιο αναλυτικά, καλύπτει άριστα τις ιατρικές επιστήμες (*medical and biomedical sciences*), σε μεγάλο βαθμό την πληροφορική και τη χημεία (*computer science and chemistry*), ικανοποιητικά τις κοινωνικές επιστήμες και την οικονομία (*social sciences and economics*),

ενώ υστερεί στην κάλυψη των μαθηματικών και των γεωλογικών επιστημών (*mathematics and earth sciences*) αλλά και της κλασσικής φιλολογίας (*classics*).

Σχετικά με τις πηγές των πληροφοριών της δεν υπάρχει καμία επίσημη ανακοίνωση, γεγονός που δημιουργεί επιφυλάξεις όσον αφορά την αξιοπιστία της. Από πειράματα έχει διαπιστωθεί πάντως ότι καλύπτει με τον καλύτερο δυνατό τρόπο το περιεχόμενο που παρέχεται δωρεάν από ακαδημαϊκές πηγές του Παγκόσμιου Ιστού (*Open Access*) ενώ εμφανίζεται να έχει συνάψει συμφωνίες με αρκετούς συνδέσμους διαφόρων επιστημονικών κλάδων (π.χ. ACM, ACS). Αξίζει επίσης να επισημανθεί ότι υποστηρίζει και μη αγλλόφωνες δημοσιεύσεις με εντυπωσιακή κάλυψη των κινέζικων δημοσιεύσεων, λόγω συμφωνιών της με δυο μεγάλες κινέζικες βάσεις δεδομένων. Όπως και οι ανταγωνιστές της, είναι όμως ισχυρότερη στα αγγλικά.

Μεγάλο της πλεονέκτημα είναι και η εξαιρετική ταχύτητά, καθώς οι αποκρίσεις της είναι ακαριαίες, όντας η ταχύτερη και με διαφορά από τις προηγούμενες υπηρεσίες.

Όσον αφορά στην ευελιξία που παρέχει στο χρήστη, διαθέτει δυνατότητες απλής αλλά και σύνθετης αναζήτησης με αρκετές επιλογές για παραμετροποίησή της (π.χ. boolean τελεστές). Υστερεί ωστόσο στην παραμετροποίηση των εμφανιζόμενων αποτελεσμάτων, ένα πολύ σημαντικό χαρακτηριστικό των ανταγωνιστών της που βελτιώνει την ευχρηστία τους. Για την ακρίβεια, τα αποτελέσματα εμφανίζονται ως επί το πλείστον ταξινομημένα με βάση το πλήθος των βιβλιογραφικών αναφορών τους, χωρίς όμως να παρέχεται η δυνατότητα εμφάνισης μόνο συγκεκριμένων τύπων εγγράφων ή χρονολογικής ταξινόμησης. Αρνητικό είναι επίσης το γεγονός ότι οι επιστρεφόμενες εγγραφές είναι σχετικά λιτές. Περιέχουν δηλαδή λιγότερες πληροφορίες από τις προηγούμενες μηχανές αναζήτησης, οι οποίες παρέχουν μακράν πιο περιεκτικές σε πληροφορίες εγγραφές.

Διαθέτει μια αρκετά καλή υπηρεσία αναζήτησης βιβλιογραφικών αναφορών, η οποία όμως δέχεται συχνά επικρίσεις για την ακρίβεια και την εγκυρότητα των αποτελεσμάτων της. Ωστόσο, ιδίως για δημοσιεύσεις από το 1990 και έπειτα, η υπηρεσία αυτή είναι συγκρίσιμη με την αντίστοιχη του WoS ([SP06]). Και αυτό γιατί εντοπίζει βιβλιογραφικές αναφορές από την επονομαζόμενη "γκρίζα βιβλιογραφία" (*grey literature*), δηλαδή τη μη δημοσιευμένη όπως αναφορές (reports), βιβλία και πρακτικά κάποιων συνεδρίων (conference proceedings). Αναμενόμενο είναι επομένως το γεγονός ότι *επιστρέφει κατά μέσο όρο περισσότερες βιβλιογραφικές αναφορές για κάθε δημοσίευση, με μικρή μάλιστα επικάλυψη με τα αποτελέσματα τόσο του Scopus όσο και του WoS*. Πρέπει να σημειωθεί ότι σε αυτό συνεισφέρει και η μεγαλύτερη υποστήριξη άλλων γλωσσών πέραν των αγγλικών.

Συνολικά, η υπηρεσία αυτή είναι αρκετά αξιόλογη με μεγάλα περιθώρια βελτίωσης των αποτελεσμάτων της αλλά και εμπλουτισμού της με επιπλέον χαρακτηριστικά που οι ανταγωνιστές της διαθέτουν ήδη. Είναι επίσης και η μόνη που προσφέρεται δωρεάν.

2.1.4 Λοιπές αξιόλογες ακαδημαϊκές μηχανές αναζήτησης

Εκτός από τις προηγούμενες μηχανές αναζήτησης, υπάρχει έντονη δραστηριότητα για την ανάπτυξη νέων μηχανών, οι οποίες είτε θα είναι άμεσα ανταγωνιστικές είτε θα επικεντρώνονται σε ένα συγκεκριμένο επιστημονικό κλάδο. Στη συνέχεια αναφέρουμε τις πιο αξιόλογες από αυτές τις προσπάθειες.

Στους άμεσους ανταγωνιστές ανήκει η μηχανή αναζήτησης **Live Search Academic**, δημιουργός της οποίας είναι η Microsoft. Αποτελεί μια εξαιρετικά καινούρια (ξεκίνησε τη λειτουργία της μόλις στις αρχές του 2007) αλλά φιλόδοξη προσπάθεια, με καλές προοπτικές εξέλιξης. Όσον αφορά τα τεχνικά χαρακτηριστικά της ([Godi07]), έχει περιορισμένες πηγές πληροφοριών (αντλεί δεδομένα μόνο από δημοσιεύσεις σε επιστημονικά περιοδικά), δεν υποστηρίζει ακόμα μη αγλλόφωνες δημοσιεύσεις, δεν προσφέρει πολλές δυνατότητες παραμετροποίησης ενώ τέλος δεν υποστηρίζει ανάλυση βιβλιογραφικών αναφορών. Οι ελλείψεις αυτές είναι εύλογες για μια τόσο καινούρια υπηρεσία αλλά αναμένεται να επιλυθούν με το πέρασμα του χρόνου.

Όσον αφορά τις εξειδικευμένες μηχανές αναζήτησης, παραθέτουμε ενδεικτικά μερικές που πιστεύουμε ότι ξεχωρίζουν:

Citeseer – http://citeseer.ist.psu.edu/ ([GBL98]): αποτελεί μια από τις πρωτοπόρες και εξαιρετικά φιλόδοξες προσπάθειες δημιουργίας ακαδημαϊκών μηχανών αναζήτησης με ποικίλες δυνατότητες, όπως ανάλυση βιβλιογραφικών αναφορών. Δημιουργήθηκε από μια ομάδα ερευνητών της NEC και ξεκίνησε τη λειτουργία της το 1997. Κεντρικό ρόλο στην όλη προσπάθεια διαδραματίζει ένα σύστημα ανάκτησης πληροφορίας (*web crawler*) που εντοπίζει *αυτόματα* όποια νέα δημοσίευση αναρτάται σε δικτυακό τόπο με ελεύθερη πρόσβαση, όπως homepage συγγραφέων. Μετά από κάποια επεξεργασία των νέων δημοσιεύσεων και την καταχώρηση τους στην αντίστοιχη βάση δεδομένων, εξετάζεται με τη βοήθεια ενός αλγορίθμου μηχανικής μάθησης αν ταυτίζονται με κάποιες από τις ήδη καταχωρημένες δημοσιεύσεις. Επιχειρείται δηλαδή να λυθεί το citation matching πρόβλημα. Μετά από σχεδόν μια δεκαετία λειτουργίας μπορούμε με ασφάλεια να πούμε ότι το εγχείρημα αυτό στέφθηκε από σημαντική επιτυχία, ωστόσο έχει ξεπεραστεί από τους ανταγωνιστές του (π.χ. Google Scholar). Επιπλέον, δεν έχει καταφέρει να επιλύσει αποτελεσματικά τα προβλήματα mixed και split citation. Για το λόγο αυτό άλλωστε συναντάται συχνά στη βιβλιογραφία ως πεδίο αξιολόγησης για αλγορίθμους επίλυσης των προβλημάτων αυτών.

DBLP (*Digital Bibliography & Library Project*) – http://dblp.uni-trier.de/ : το περιεχόμενό της αφορά αποκλειστικά της επιστήμη των υπολογιστών και καλύπτει τα σημαντικότερα περιοδικά και συνέδρια του τομέα. Συνολικά περιλαμβάνει 90000 περίπου έγγραφες και

μερικές χιλιάδες συνδέσμους στις προσωπικές ιστοσελίδες επιστημόνων του τομέα. Δεν υποστηρίζει όμως ανάλυση βιβλιογραφικών αναφορών. Εκτός από ενημερωτικούς λόγους, χρησιμοποιείται συχνά για τη δημιουργία datasets με τα οποία θα εκπαιδευτούν αλγόριθμοι μηχανικής μάθησης. Σε αυτό συντελεί και το γεγονός ότι το σύνολο των πληροφοριών της είναι κωδικοποιημένο σε μορφή xml.

Τέλος, υπάρχει πλήθος άλλων σημαντικών εξειδικευμένων ακαδημαϊκών μηχανών αναζήτησης με ποικίλες δυνατότητες, όπως η **ACM Digital Library** για την επιστήμη των υπολογιστών, το **PudMed Central** για την ιατρική και η **SciFinder Scholar** που εξειδικεύεται στη χημεία. Η απαρίθμησή τους είναι ωστόσο εκτός του σκοπού μας και για περισσότερες πληροφορίες παραπέμπουμε στο [BMR06], σελ. 15.

2.1.5 Συμπεράσματα για τις ακαδημαϊκές μηχανές αναζήτησης

Συμπερασματικά, μπορούμε να ισχυριστούμε ότι ενώ οι τρεις κυριότερες μηχανές αναζήτησης είναι φαινομενικά ανταγωνιστικές, στην πραγματικότητα *συμπληρώνουν η μια την άλλη*. Και αυτό γιατί καμία δεν θεωρείται ακόμα πλήρης. Αντίθετα, κάθε μια, λόγω των ιδιαιτεροτήτων της (διαφορετικών πηγών πληροφοριών, διαφορετικών μεθοδολογιών επεξεργασίας τους αλλά και υποστηριζόμενων γλωσσών), εξειδικεύεται και πλεονεκτεί σε διαφορετικό τομέα από τις υπόλοιπες. Για παράδειγμα, το Scopus αποτελεί την ιδανική επιλογή για την αναζήτηση βιβλιογραφίας πιο πρόσφατης από το 1995 στον τομέα των ιατρικών επιστημών. Στην άποψη αυτή (της συμπληρωματικότητας) συνηγορεί άλλωστε και η συνήθως μικρή σχετικά επικάλυψη των αποτελεσμάτων τους.

Πρέπει επίσης να τονιστεί το γεγονός ότι ο υψηλός μεταξύ τους ανταγωνισμός επιβάλλει τη συνεχή βελτίωση των αποτελεσμάτων τους αλλά και εξέλιξή τους για την προσθήκη νέων δυνατοτήτων και χαρακτηριστικών. Χαρακτηριστική είναι η εντατική έρευνα που βρίσκεται σε εξέλιξη από τις Scopus και WoS για να τεθεί σε λειτουργία μια νέα δυνατότητα που ονομάζεται Ταυτοποίηση Συγγραφέα (*Author Identification*), η οποία στην ουσία αφορά την επίλυση των προβλημάτων mixed και split citation στις βάσεις δεδομένων τους. Επόμενο είναι οι συσχετίσεις μεταξύ των κυρίαρχων αυτών μηχανών αναζήτησης να αλλάζουν συνεχώς.

Συνεπώς, θα ήταν επιθυμητό να μπορούσαμε να συμπεριλάβουμε στην προσπάθειά μας πληροφορίες και από τις τρεις αυτές μηχανές αναζήτησης. Ωστόσο, λόγω του διόλου ευκαταφρόνητου όγκου εργασίας που θα απαιτούσε κάτι τέτοιο περιοριστήκαμε σε μια από αυτές, χωρίς να αποκλείεται μελλοντική υποστήριξη και των υπολοίπων. Επιλέξαμε το Google Scholar με το σκεπτικό ότι αφενός επιστρέφει αναλογικά περισσότερα citations για

κάθε δημοσίευση και αφετέρου έχει τις καλύτερες προοπτικές βελτίωσης. Επιπλέον, δεν απαιτεί συνδρομή ενώ είναι και πιο εύκολο το parsing των αποτελεσμάτων του.

2.2 Αυτοματοποίηση της συλλογής και ταύτισης βιβλιογραφικών εγγραφών

2.2.1 Βασικές έννοιες μηχανικής μάθησης

Η ενότητα αυτή είναι αφιερωμένη στις πιο σημαντικές προσπάθειες επίλυσης των προβλημάτων mixed citation, split citation, και citation matching, οι οποίες στηρίζονται σε μεθόδους μηχανικής μάθησης.

Η **Μηχανική Μάθηση** (*machine learning*) αποτελεί ένα παρακλάδι της Τεχνητής Νοημοσύνης (*artificial intelligence*) που έχει ως αντικείμενο την ανάπτυξη τεχνικών και μεθόδων που επιλύουν *προσεγγιστικά αλλά με ικανοποιητική ακρίβεια* προβλήματα για τα οποία δεν έχει ακόμα επινοηθεί ή δεν είναι δυνατόν να βρεθεί *ακριβής* μαθηματική-αλγοριθμική λύση. Τα αντίστοιχα προγράμματα λέμε ότι μαθαίνουν από την εμπειρία τους **E** ως προς ένα σύνολο εργασιών **T** με μέτρο απόδοσης **P** εννοώντας ότι η απόδοσή τους στις εργασίες του **T** εκτιμώμενες από το μέτρο **P** βελτιώνεται με την εμπειρία **E**. ([Mit97]).

Η μάθηση αυτή διακρίνεται περαιτέρω στις εξής βασικές κατηγορίες:

- **επιβλεπόμενη μάθηση** (*supervised learning*), κατά την οποία παρουσιάζεται στο πρόγραμμα ένα σύνολο δεδομένων εκπαίδευσης της μορφής <διάνυσμα εισόδου, τιμή εξόδου>. Στόχος είναι να εκπαιδευθεί το πρόγραμμα ώστε να μάθει μια συνάρτηση η οποία προσεγγίζει όσο γίνεται καλύτερα τα δεδομένα εισόδου αλλά ταυτόχρονα είναι αρκετά γενική ώστε να εφαρμόζεται επιτυχώς και σε άγνωστα δεδομένα. Πρόκειται επομένως για επαγωγική μάθηση, αφού το ζητούμενο είναι να καταλήξει το πρόγραμμα σε μια υπόθεση η οποία θα του εξασφαλίσει ικανότητα σωστής απόφασης σε περιπτώσεις που δεν έχει ξαναδεί. Ειδική μέριμνα λαμβάνεται ωστόσο για να μην καταλήξει σε προφανείς (trivial) υποθέσεις. Αυτό εξασφαλίζεται με την λεγόμενη επαγωγική κλίση (*inductive bias*) που χαρακτηρίζει τους αλγορίθμους της μηχανικής μάθησης.
- **μη επιβλεπόμενη μάθηση** (*unsupervised learning*), κατά την οποία παρουσιάζεται στο πρόγραμμα ένα σύνολο δεδομένων χωρίς επιθυμητές τιμές εξόδου. Στόχος είναι να εντοπίσει το πρόγραμμα την εσωτερική δομή των δεδομένων αυτών.

Τυπικά πεδία εφαρμογής των παραπάνω μεθόδων μάθησης αποτελούν *αντίστοιχα*:

- Τα **προβλήματα ταξινόμησης** (*classification*). Στα προβλήματα αυτά η τιμή εξόδου είναι διακριτή και αντιπροσωπεύει την ομάδα-κλάση στην οποία ανήκει το διάνυσμα εισόδου. Το πρόγραμμα εκπαιδεύεται επομένως ώστε να είναι σε θέση να κατατάσσει τα διανύσματα εισόδου στην κλάση στην οποία ανήκουν. Ευρέως διαδεδομένοι ταξινομητές (*classifiers*) είναι τα Support Vector Machines (*SVMs*) και ο Naive Bayes Classifier. Για περισσότερες πληροφοριες ο αναγνώστης μπορεί να ανατρέξει στο [Mit97] και το [Mit06].
- Τα **προβλήματα ομαδοποίησης** (*clustering*). Στα προβλήματα αυτά στόχος είναι να διαχωριστεί το σύνολο των διανυσμάτων εισόδου σε υποσύνολα (*clusters*), συνήθως ξένα μεταξύ τους, τέτοια ώστε τα στοιχεία τους να έχουν όσο το δυνατό πιο όμοια δομή. Ταυτόχρονα τα clusters πρέπει να είναι όσο το δυνατό πιο ανόμοια μεταξύ τους. Μια καλή επισκόπηση των αλγορίθμων ομαδοποίησης (από την πλευρά του Data Mining που μας ενδιαφέρει περισσότερο στην εργασία αυτή) γίνεται στο [GR02].

Όπως θα φανεί και στις περιπτώσεις των προβλημάτων mixed και split citation, συχνά είναι δύσκολο να αποφασιστεί σε ποια από τις δυο κατηγορίες κατατάσσεται ένα πρόβλημα ώστε να εφαρμοστεί η αντίστοιχη μέθοδος μάθησης για την επίλυσή τους.

2.2.2 Mixed & Split Citation Problem

Στην ενότητα αυτή επιχειρούμε να καταγράψουμε τη σημαντικότερη βιβλιογραφία σχετικά και με τα δυο αυτά ζητήματα. Και αυτό γιατί, παρ' όλο που αποτελούν διαφορετικά προβλήματα, συχνά συγχέονται και δεν τονίζεται η διαφορά τους.

Τυπικά τα προβλήματα αυτά ορίζονται ως εξής:

Mixed Citation Problem ([LOPK05])**:**

Δεδομένου ενός συνόλου βιβλιογραφικών αναφορών, έστω ***C****, που αφορούν ένα συγγραφέα* $\boldsymbol{a_i}$*, να εντοπιστούν γρήγορα και με ακρίβεια οι βιβλιογραφικές αναφορές που στην πραγματικότητα είναι γραμμένες από ένα διαφορετικό συγγραφέα* $\boldsymbol{a_j}$*, που τυχαίνει να έχει το ακριβώς ίδιο όνομα.*

Το πρόβλημα αυτό είναι ιδιαίτερα έντονο σε περιπτώσεις συνήθων επωνύμων. Ένα χαρακτηριστικό παράδειγμα ([HGZ04]) εντοπίζεται στο Citeseer αναζητώντας τον επιστήμονα με τις περισσότερες βιβλιογραφικές αναφορές στον κλάδο της επιστήμης των υπολογιστών. Ο επιστήμονας αυτός εμφανίζεται να είναι ο D. Johnson, ο οποίος στην πραγματικότητα αποτελεί συνδυασμό διαφόρων επιστημόνων, όπως David B. Johnson, David S. Johnson ακόμα και Joel T. Johnson κλπ.

Split Citation Problem ([OLK05]):

Δεδομένων δυο λιστών με ονόματα συγγραφέων, έστω $\boldsymbol{X}$ *και* $\boldsymbol{Y}$*, για κάθε όνομα συγγραφέα* $\boldsymbol{x}$ $(\boldsymbol{\in X})$*, να βρεθεί ένα σύνολο ονομάτων,* $\boldsymbol{y_1, y_2, \ldots, y_n}$ $(\boldsymbol{\in Y})$*, τέτοια ώστε τόσο το* $\boldsymbol{x}$ *όσο και τα* $\boldsymbol{y_i}$ $(\boldsymbol{1 \le i \le n})$ *να αποτελούν παραλλαγές του ονόματος του ίδιου συγγραφέα.*

Ένα χαρακτηριστικό παράδειγμα για την περίπτωση αυτή ([LOPK05]) εντοπίζεται στην ACM Digital Library, όπου ο φημισμένος επιστήμονας πληροφορικής Jeffrey D. Ullman εμφανίζεται με 10 διαφορετικά ονόματα. Παραθέτουμε ενδεικτικά ορισμένα από αυτά: J. Ullman, J. D. Ullman, Jeff Ullman, Jeffrey D. Ullman, Jeffrey Ullman, J. Ullmann, Jeffrey D. Ullmann.

Πρέπει να επισημανθεί ωστόσο ότι οι όροι αυτοί δεν καθολικά αποδεκτοί από όλους τους ερευνητές του τομέα. Αντίθετα, τα προβλήματα αυτά εμφανίζονται με μια πληθώρα διαφορετικών ονομασιών στη βιβλιογραφία. Πιο συχνά χρησιμοποιείται ο όρος **name disambiguation**, συνδυάζοντας τα προβλήματα mixed και split citation, ενώ με την ίδια σημασία χρησιμοποιείται και ο όρος **name equivalence identification**. Τέλος, σε ορισμένες εργασίες αναφέρεται ο όρος **name matching**, υπονοώντας το split citation problem (μόνο).

Τα προβλήματα αυτά εντάσσονται μαζί με το citation matching στο γενικότερο πρόβλημα **identity uncertainty**, το οποίο επίσης συναντάται με διάφορους όρους στη σχετική βιβλιογραφία: record linkage, object identity, object matching, merge-purge, duplicate detection κλπ. Το ζητούμενο στο πρόβλημα αυτό είναι να εντοπιστούν τα διαφορετικά αντικείμενα που αντιστοιχούν στην ίδια οντότητα. Επομένως, το πρόβλημα είναι γενικότερο και δεν συναντάται μόνο στο πλαίσιο των βιβλιογραφικών αναφορών.

Όσον αφορά τα <u>αίτια των προβλημάτων αυτών</u>, μεταξύ αυτών συγκαταλέγονται: ορθογραφικά λάθη, χρήση υποκοριστικών ονομάτων (π.χ. Jeff αντί Jeffrey), λάθη στο λογισμικό συλλογής βιβλιογραφικών αναφορών, αλλαγή ονομάτων λόγω γάμου, αντιμετάθεση ονομάτων και επιθέτων (π.χ. Han Hui αντί Hui Han) αλλά και στο γεγονός ότι ορισμένα ονόματα και επίθετα είναι πολύ συνηθισμένα. Επίσης, το πρόβλημα επιτείνεται από το γεγονός ότι δεν υπάρχει ένας καθιερωμένος τρόπος αναγραφής των ονομάτων στις βιβλιογραφικές αναφορές. Δηλαδή σε ορισμένες αναφέρονται οι συγγραφείς με πλήρες το μικρό τους όνομα και σε άλλες μόνο με το αρχικό του γράμμα. Εμείς θα θεωρήσουμε στα πλαίσια της εργασίας αυτής ότι συγγραφείς δηλώνονται μόνο με το πρώτο γράμμα του ονόματός τους ακολουθούμενο από το επίθετο τους. Αυτός άλλωστε είναι και ο συνηθέστερος τρόπος.

Δυο είναι οι σημαντικότερες ομάδες που ασχολούνται με τα προβλήματα mixed και split citaiton, αμφότερες με έδρα το Pennsylvania State University.

Η πρώτη από αυτές, αποτελούμενη κυρίως από τους **H. Han, H. Zha και C. Giles**, δεν διαχωρίζει ρητά τα δυο προβλήματα, αλλά αναφέρεται από κοινού σε αυτά χρησιμοποιώντας τον όρο *name disambiguation in author citation*. Θεωρητικά επομένως, οι μέθοδοι που προτείνουν στοχεύουν στην ταυτόχρονη επίλυση και των δυο προβλημάτων, στην πράξη όμως επικεντρώνονται πολύ περισσότερο στο mixed citation.

Η ομάδα αυτή έχει προσεγγίσει το ζήτημα τόσο από την οπτική της επιβλεπόμενης μάθησης-classification ([HGZ04]) όσο και της μη επιβλεπόμενης-clustering ([HXZG05], [HZG05]), καταλήγοντας τελικά στο ότι η δεύτερη είναι η βέλτιστη προσέγγιση. Το σημαντικότερο *πλεονέκτημα της μη επιβλεπόμενης μάθησης* είναι το γεγονός ότι δεν απαιτεί επεξεργασμένα σύνολα δεδομένων για εκπαίδευση. Κοινός άξονας σε όλες τις προσεγγίσεις τους είναι οι ιδιότητες των βιβλιογραφικών αναφορών που χρησιμοποιούνται για την ταυτοποίησή τους (*ονόματα συν-συγγραφέων, τίτλος της δημοσίευσης, τίτλος του περιοδικού ή συνεδρίου στο οποίο έγινε η δημοσίευση*). Κοινό επίσης παραμένει σε όλες τις περιπτώσεις το είδος των δεδομένων με τα οποία διεξάγονται όλα τα πειράματα (datasets από το DBLP και από προσωπικές ιστοσελίδες των συγγραφέων). Αυτό καθιστά άμεσα συγκρίσιμα τα αποτελέσματα των διαφορετικών προσεγγίσεων.

Γενικά, τα κυριότερα συμπεράσματα στα οποία καταλήγουν μετά από μια σειρά ποικίλων πειραμάτων στις (τρεις) δημοσιεύσεις τους είναι τα εξής:

- Δεν συνεισφέρουν και οι τρεις χρησιμοποιούμενες ιδιότητες το ίδιο στην ακρίβεια μιας απόφασης. Αναλυτικότερα, τα ονόματα των συν-συγγραφέων αποτελούν την πιο σθεναρή ιδιότητα, ικανή από μόνη της να επιτύχει υψηλή ακρίβεια. Ακολουθούν οι τίτλοι των περιοδικών/συνεδρίων ενώ τη μικρότερη συνεισφορά έχει ο τίτλος της δημοσίευσης. Μάλιστα, η συνεισφορά των δυο τελευταίων ιδιοτήτων βελτιώνεται αισθητά όταν δεν απαιτείται οι αντίστοιχες ιδιότητες δυο δημοσιεύσεων να περιέχουν ακριβώς ίδιες λέξεις, αλλά αντίθετα επιδιώκεται *σημασιολογική ταύτιση*. Αυτό υπαγορεύει άλλωστε η κοινή λογική, επειδή αφενός ένας συγγραφέας σπάνια χρησιμοποιεί ακριβώς τις ίδιες λέξεις στον τίτλο των δημοσιεύσεών του και αφετέρου δεν τις δημοσιεύει συνέχεια στα ίδια περιοδικά ή συνέδρια. Αντίθετα, είναι πολύ πιθανότερο να χρησιμοποιήσει διαφορετικές λέξεις με ίδια ωστόσο σημασία, λέξεις δηλαδή ενδεικτικές του τομέα στον οποίο εξειδικεύεται. Η σημασιολογική ταύτιση εφαρμόστηκε στο **[HXZG05]**, όπου χρησιμοποιήθηκε μια μέθοδος ομαδοποίησης των λέξεων των τίτλων με την ίδια σημασία στο ίδιο σύνολο. Όπως ήταν αναμενόμενο, η ακρίβεια βελτιώθηκε σημαντικά. Επιπλέον, στο **[HZG05]** εξετάζεται το θέμα απονομής βαρών σε κάθε μια από τις ιδιότητες αυτές. Υποστηρίζουν ότι ενδεχομένως τα αποτελέσματα να βελτιώνονταν αν ένας αλγόριθμος επιβλεπόμενης μάθησης αναλάμβανε αύτη την απονομή βαρών.

- Η διαφορά μεταξύ της επίδοσης των δυο classifiers που προτάθηκαν στο **[HGZ04]** (Naive Bayes-generative classifier & SVM-discriminative classifier) είναι ελάχιστη. Αυτό οφείλεται στα διαφορετικά πλεονεκτήματα που προσφέρει ο κάθε ένας. Πιο συγκεκριμένα, ο Naive Bayes Classifier επιτυγχάνει καλή απόδοση στην αποκάλυψη των μοτίβων συνεργασιών ενός συγγραφέα. Από την άλλη μεριά, η απόδοση του SVM εξασφαλίζεται από το γεγονός ότι αποκαλύπτει καλύτερα τα χαρακτηριστικά που είναι μοναδικά σε κάθε συγγραφέα.
- Όσον αφορά στη μη επιβλεπόμενη μάθηση, προτείνονται δυο διαφορετικοί μέθοδοι: μια σύνθετη μέθοδος πιθανοτικής ομαδοποίησης (**[HXZG05]**) και ένας φασματικός (*spectral*) αλγόριθμος ομαδοποίησης (**[HZG05]**). Και οι δυο μέθοδοι συγκρινόμενες με τον διαδεδομένο αλγόριθμο ομαδοποίησης K-means επιτυγχάνουν καλύτερα αποτελέσματα, ενώ γενικότερα την καλύτερη απόδοση σημειώνει η δεύτερη μέθοδος. Ανοικτό μένει ωστόσο και στις δύο μεθόδους το πρόβλημα του αυτόματου καθορισμού του αριθμού των clusters, το οποίο στα πειράματα που διεξήχθησαν ήταν καθορισμένο εξ' αρχής.

Η δεύτερη ομάδα, με κύριους συντελεστές τους B. On, D. Lee και J. Kang, θίγει εξίσου και τα δύο προβλήματα αλλά επικεντρώνεται περισσότερο στην κλιμακωσιμότητα (scalability) των μεθόδων που τα επιλύουν. Δηλαδή οι προσεγγίσεις της είναι περισσότερο προσανατολισμένες στην επεξεργασία τεράστιων σύνολων δεδομένων. Πιο συγκεκριμένα, υποστηρίζεται ότι η επιθυμητή ιδιότητα της κλιμακωσιμότητας επιτυγχάνεται με τη βοήθεια ενός σταδίου φιλτραρίσματος (blocking) που προηγείται της κύριας επεξεργασίας. Έτσι, μειώνονται στο ελάχιστο οι συγκρίσεις που αυτή θα πραγματοποιήσει. Οι αλγόριθμοι που προτείνονται είναι γραμμικής σχεδόν πολυπλοκότητας χάρη στα δυο στάδια που περιλαμβάνουν: το πρώτο αποκλείει τις προφανώς άχρηστες συγκρίσεις, ενώ το δεύτερο υλοποιεί την κυρίως επεξεργασία για την επίλυση του προβλήματος. Πρέπει επίσης να σημειωθεί ότι και στις δυο δημοσιεύσεις τους χρησιμοποιούν τις *ίδιες ιδιότητες με την πρώτη ομάδα* (λίστα συν-συγγραφέων, τίτλος δημοσίευσης, τίτλος περιοδικού/συνεδρίου). Ωστόσο δεν είναι δυνατό να συγκριθούν άμεσα οι μέθοδοί τους καθώς χρησιμοποιούν διαφορετικά σύνολα δεδομένων, εστιάζουν σε διαφορετικά θέματα (π.χ. scalability) και εξετάζουν την επίδραση διαφορετικών παραγόντων στην επιτυχία των πειραμάτων τους.

Αναλυτικότερα, στο **[OLK05]** επικεντρώνονται στο πρόβλημα split citation, για το οποίο εξετάζουν διάφορες μεθόδους για κάθε ένα από τα δυο στάδια. Ενδιαφέρον είναι το γεγονός ότι για το στάδιο της κυρίως επεξεργασίας εξετάζουν μεθόδους τόσο επιβλεπόμενης (Naive Bayes & SVM) όσο και μη-επιβλεπόμενης (string distance metrics – ενότητα 2.2.4), καταλήγοντας στην ανωτερότητα των δεύτερων. Για την ακρίβεια καταλήγουν σε συμπεράσματα παρόμοια με το [CRF03] που παρουσιάζουμε στη συνέχεια. Στο **[LOPK05]**

εξετάζουν τόσο το mixed όσο και το split citation πρόβλημα. Για το πρώτο όμως παρουσιάζουν έναν απλοϊκό αλγόριθμο, ενώ για το δεύτερο επαναλαμβάνουν την μέθοδο της [OLK05].

Τέλος, στο **[MSS06]** προτείνεται μια μέθοδος για την επίλυση του name disambiguation που διαφέρει ριζικά από τις άλλες, επιτυγχάνοντας εξίσου ικανοποιητικά αποτελέσματα. Πιο συγκεκριμένα, η μέθοδος αυτή βασίζεται στο γεγονός ότι *οι επιστήμονες έχουν την τάση να αναφέρουν στη βιβλιογραφία των δημοσιεύσεών τους προηγούμενες εργασίες τους* (**self-citation**). Συνδυάζοντας τα self-citations με τους συν-συγγραφείς και τα URLs των papers η προτεινόμενη μέθοδος ομαδοποιεί ένα σύνολο δημοσιεύσεων (για την ακρίβεια διαμερίζει τον αντίστοιχο γράφο) έτσι ώστε κάθε ομάδα να περιέχει papers του ίδιου συγγραφέα. Η προσέγγιση αυτή είναι περισσότερο κοντά στη δική μας από την άποψη ότι είναι εξ' αρχής προσανατολισμένη στην επεξεργασία αποτελεσμάτων ακαδημαϊκών μηχανών αναζήτησης. Αντίθετα, όλες οι προηγούμενες προσπάθειες εφαρμόζονταν σε ένα προκαθορισμένο σύνολο βιβλιογραφικών αναφορών. Γι' αυτό άλλωστε η μέτρηση της απόδοσης της μεθόδου της [MSS06] βασίζεται σε μετρικές της ανάκτησης πληροφοριών (precision, recall, f-measure) και όχι στην ακρίβεια (accuracy) επί του dataset.

Σε όλες τις περιπτώσεις που αναφέραμε επισημαίνεται ότι είναι καλό να χρησιμοποιηθούν περισσότερες ιδιότητες-χαρακτηριστικά για να αυξηθεί η ακρίβεια. Σε αυτό το θέμα εστιάζουν οι εργασίες [KMP06] και [TKL06]. Αναλυτικότερα, σε αυτές προτείνονται διάφορες μέθοδοι για ανάκτηση συμπληρωματικών πληροφοριών από το Διαδίκτυο για τη σχέση μεταξύ δυο βιβλιογραφικών αναφορών με σκοπό την ορθότερη επίλυση του mixed citation (μόνο).

Πιο συγκεκριμένα, η **[KMP06]** μοντελοποιεί το πρόβλημα ως διαμερισμό ενός *γράφου* που έχει ως *κορυφές* βιβλιογραφικές αναφορές που αντιστοιχούν σε ένα συγγραφέα και οι *ακμές* του έχουν βάρη ανάλογα με τη συσχέτιση των αντίστοιχων κορυφών. Δυο κύριες μέθοδοι προτείνονται με κοινό στοιχείο το ότι χρησιμοποιούν μια μηχανή αναζήτησης (Google) για να κάνουν ένα ερώτημα που περιέχει τους τίτλους των δημοσιεύσεων των δυο κορυφών. Η πρώτη χρησιμοποιεί τα αποτελέσματα για να επηρεάσει το βάρος της ακμής που ορίζουν οι δυο κορυφές, ενώ η δεύτερη προσθέτει νέους κόμβους για τις πρώτες από τις επιστρεφόμενες ιστοσελίδες. Στη συνέχεια, επανεξετάζει με βάση τη μεταβατική ιδιότητα τις συσχετίσεις των αρχικών κορυφών. Συνδέει δηλαδή μεταξύ τους όσες από τις αρχικές κορυφές είναι όμοιες με την ίδια νέα κορυφή.

Όσον αφορά το **[TKL06]**, η πληροφορία που αντλείται από τον Παγκόσμιο Ιστό αξιοποιείται από ένα αλγόριθμο ομαδοποίησης (HAC) ως εξής: Σε κάθε δημοσίευση $\mathbf{c \in C}$ αποδίδεται ένα σύνολο σχετικών URL που επιστρέφει μια μηχανή αναζήτησης. *Κάθε URL έχει ένα βάρος*

αντιστρόφως ανάλογο της συχνότητας της συγκεκριμένης ιστοσελίδας, με το σκεπτικό ότι οι σπάνιες ιστοσελίδες μας οδηγούν σε πιο σίγουρα συμπεράσματα. Στη συνέχεια, ομαδοποιούνται οι δημοσιεύσεις του **C** ανάλογα με την ομοιότητα των URL διευθύνσεών τους έτσι ώστε τελικά να καταλήξουμε σε έναν αριθμό clusters που το καθένα από αυτά αντιστοιχεί σε ένα συγγραφέα. Αν και η μέθοδος αυτή δεν επιτυγχάνει μεγάλη ακρίβεια (μόλις 0.836 κατά μέσο όρο), πρέπει να αναλογιστεί κανείς ότι χρησιμοποιεί μόνο μια ιδιότητα, σε αντίθεση με τις προηγούμενες προσπάθειες που αξιοποιούσαν τρεις ιδιότητες. Εκτός αυτού, τα συμπεράσματα της είναι χρήσιμά για την περίπτωσή μας, αφού μια από τις ιδιότητες που μας παρέχει το Google Scholar είναι και οι URLs διευθύνσεις των δημοσιεύσεων.

Στο σημείο αυτό πρέπει να επισημάνουμε το γεγονός ότι στο σύνολο των προαναφερθέντων προσπαθειών επιτυγχάνονται φτωχά αποτελέσματα στην περίπτωση που ένας συγγραφέας έχει μικρό αριθμό δημοσιεύσεων αλλά και στις περιπτώσεις που για κάποιες εθνότητες (π.χ. Κορεάτες) το πλήθος των διαφορετικών επιθέτων είναι περιορισμένο. Αντίθετα, καλύτερα αποτελέσματα επιτυγχάνονται όταν οι συγγραφείς με τα ίδια ή παρόμοια ονόματα εξειδικεύονται σε διαφορετικές επιστημονικούς τομείς.
Τέλος, αξίζει να αναφερθεί ότι όλοι οι ερευνητές προτείνουν ως βέλτιστη λύση για τα προβλήματα mixed και split citation την απόδοση ενός μοναδικού αναγνωριστικού κωδικού (identifier) σε κάθε συγγραφέα στα πρότυπα αντίστοιχων επιτυχημένων προσπαθειών. Ως παραδείγματα αναφέρονται η λίστα ονομάτων καλλιτεχνών του μουσείου Getty (*Getty's Union List of Artist's Names*) αλλά και το αρχείο ονομάτων της βιβλιοθήκης του αμερικανικού Κογκρέσου (*Library of Congress name authoritative file*).
Η μέθοδος που προτείνουμε για την επίλυση των προβλημάτων αυτών παρουσιάζεται στο κεφάλαιο 3.

2.2.3 Citation Matching Problem

Στην ενότητα αυτή θα αναφερθούμε *ακροθιγώς* στη σημαντικότερη βιβλιογραφία που υπάρχει για το πρόβλημα αυτό, κυρίως γιατί στην περίπτωσή μας δεν το αντιμετωπίζουμε στη συνήθη του μορφή.
Αν και το citation matching δεν έχει οριστεί τυπικά σε καμία από τις σχετικές δημοσιεύσεις, θα μπορούσαμε να το διατυπώσουμε (στα πρότυπα του mixed citation) ως εξής:

Citation Matching Problem:

*Δεδομένου ενός συνόλου βιβλιογραφικών αναφορών, έστω **C**, να εντοπιστούν γρήγορα και με ακρίβεια οι βιβλιογραφικές αναφορές που στην πραγματικότητα αφορούν την ίδια δημοσίευση.*

Ένα παράδειγμα ενδεικτικό του προβλήματος είναι το ακόλουθο ([PMMR03]):

[Lashkari et al 94] Collaborative Interface Agents, Yezdi Lashkari, Max Metral, and Pattie Maes, Proceedings of the Twelfth National Conference on Articial Intelligence, MIT Press, Cambridge, MA, 1994.

Metral M. Lashkari, Y. and P. Maes. Collaborative interface agents. In Conference of the American Association for Artificial Intelligence, Seattle, WA, August 1994.

Οι δύο αυτές βιβλιογραφικές αναφορές, αν και διαφέρουν σημαντικά μεταξύ τους, πιθανότατα αναφέρονται στην ίδια δημοσίευση.

Ως κυριότερες αιτίες για το πρόβλημα citation matching θεωρούνται οι εξής ([Shim]):

- Το γεγονός ότι η σειρά των πεδίων (συγγραφείς, τίτλος κλπ) δεν είναι σταθερή αλλά ποικίλει.
- Κάποια πεδία σε ορισμένες βιβλιογραφικές αναφορές παραλείπονται.
- Αντιμετάθεση των ονομάτων των συγγραφέων με τα επίθετα τους, δηλαδή αντί να δηλώνονται οι συγγραφείς στη μορφή όνομα-επίθετο εμφανίζονται ως επίθετο-όνομα.
- Παράθεση των ονομάτων συγγραφέων ή των τίτλων των περιοδικών/συνεδρίων στη συντομευμένη τους μορφή.
- Ορθογραφικά λάθη.

Πρώτη αναφορά στο πρόβλημα αυτό γίνεται στα **[GBL98]** και **[LGB99]** από τους δημιουργούς του Citeseer. Στις δημοσιεύσεις αυτές συγκρίνουν την απόδοση στο ίδιο dataset τεσσάρων διαφορετικών μεθόδων και καταλήγουν στον εξής αλγόριθμο: πρώτα κανονικοποιούνται οι βιβλιογραφικές αναφορές, στη συνέχεια ταξινομούνται με βάση το μήκος τους και τέλος εξετάζεται η ύπαρξη όμοιων λέξεων στα αντίστοιχα πεδία τους.

Στο **[PMMR03]** περιγράφεται με λεπτομέρεια ένας πιο σύνθετος αλγόριθμος που βασίζεται στα relational probability models (RPMs), και συγκεκριμένα στο Markov Chain Monte Carlo. Οι συγγραφείς χρησιμοποιούν για την αξιολόγησή του το ίδιο dataset με το [LGB99] και διαπιστώνουν ότι η απόδοσή του είναι αισθητά καλύτερη από αυτή του αλγορίθμου του Citeseer.

Τέλος, στο **[WMPH04]** προτείνεται μια μέθοδος συνδυασμού (παράλληλης εκτέλεσης) της διαδικασίας εξαγωγής πληροφορίας και μιας διαδικασίας επίλυσης του citation matching με σκοπό την εκμετάλλευση των ενδιάμεσων αποτελεσμάτων της μιας διαδικασίας από την άλλη. Η μέθοδος αυτή βασίζεται στα conditional random fields (CRFs). Μάλιστα, εφαρμοζόμενη στο ίδιο dataset με το [LGB99] αποδεικνύεται ανώτερη έναντι του προτεινόμενου στο [LGB99] αλγορίθμου.

Δεδομένου όμως ότι το σύστημά μας αντλεί τις βιβλιογραφικές αναφορές από το Google Scholar, το οποίο ήδη επιλύει σε κάποιο βαθμό το citation matching, καμία από αυτές τις μεθόδους δεν είναι άμεσα εφαρμόσιμη στην περίπτωσή μας. Άλλωστε δεν αντιμετωπίζουμε το πρόβλημα στην κανονική του μορφή. Αντίθετα, ένα μέρος των αποτελεσμάτων που επιστρέφει το Google Scholar παρουσιάζει ένα πρόβλημα *παρεμφερές με το citation matching*. Πιο συγκεκριμένα, λόγω σφαλμάτων κατά την επεξεργασία των βιβλιογραφικών αναφορών, ένα μέρος από αυτές εμφανίζονται με τα εξής προβλήματα:

- Ο τίτλος μιας δημοσίευσης δεν είναι ο πραγματικός, είτε λόγω ορθογραφικών λαθών είτε κυρίως λόγω προσθήκης (*concatenation*) στην αρχή ή στο τέλος του αυθεντικού τίτλου άλλων πληροφοριών (συνήθως ο τίτλος του συνεδρίου ή του περιοδικού). Παραθέτουμε ενδεικτικά δυο παραλλαγές του τίτλου της δημοσίευσης "*On power law relationships of the Internet topology*":

  BOn PowerYLaw Relationships of the Internet Topology

  On power law relationships of the Internet topology [C]□ Proceedings SIGCOMM'99

- Τα πεδία της βιβλιογραφικής αναφοράς περιέχουν λανθασμένη πληροφορία. Για παράδειγμα, συχνό είναι το φαινόμενο ο τίτλος της δημοσίευσης να εμφανίζεται σαν τίτλος περιοδικού ή στη θέση του τίτλου να εμφανίζεται ο τόπος διεξαγωγής του συνεδρίου. Για παράδειγμα, στην ακόλουθη εικόνα, σαν τίτλος εμφανίζεται το όνομα ενός από τους συγγραφείς, ενώ ο πραγματικός τίτλος έχει τοποθετηθεί στη θέση του συνεδρίου.

Εικόνα 1. Εσφαλμένη βιβλιογραφική εγγραφή του Google Scholar

Περισσότερες στοιχεία για το πρόβλημα και η προσέγγιση που τελικά επιλέξαμε για την επίλυσή του παρουσιάζονται στην ενότητα 3.3.

2.2.4 Μετρικές Μορφολογικής Απόστασης Συμβολοσειρών (String Distance Metrics)

Παρατηρώντας κανείς τη βιβλιογραφία για τα παραπάνω προβλήματα, διαπιστώνει ότι στο σύνολο τους οι σχετικές προσπάθειες βασίζονται στις **μετρικές μορφολογικής απόστασης συμβολοσειρών**. Με τη βοήθειά τους εξετάζεται σε κάθε περίπτωση ο βαθμός στον οποίο διαφέρουν (ή ισοδύναμα μοιάζουν) οι αντίστοιχες ιδιότητες (ονόματα συγγραφέων, τίτλοι κλπ) των εκάστοτε συγκρινόμενων αντικειμένων. Κάθε άλλη προσπάθεια επίλυσης των προβλημάτων αυτών πρέπει επομένως να γίνει με πλήρη γνώση των δυνατοτήτων αλλά και των αδυναμιών των μετρικών που προτείνονται στη βιβλιογραφία. Σημαντική πηγή

πληροφόρησης αλλά και αξιολόγησης των σημαντικότερων από τις μετρικές αυτές αποτελεί η εργασία [Ραπ07]. Ενδιαφέροντα είναι, επίσης, τόσο τα πειράματα όσο και τα συμπεράσματα στα οποία καταλήγει η εργασία [CRF03].

Σε γενικές γραμμές, οι μετρικές σύγκρισης συμβολοσειρών κατατάσσονται στις ακόλουθες κατηγορίες ανάλογα με τον τρόπο αναπαράστασης των δεδομένων που συγκρίνουν:

- **αναπαράσταση με χαρακτήρες**

  Οι μετρικές της κατηγορίας αυτής συγκρίνουν τους χαρακτήρες των συμβολοσειρών, δηλαδή η ομοιότητα/διαφορά των συμβολοσειρών ορίζεται *σε επίπεδο χαρακτήρα*. Αυτή η προσέγγιση τους δίνει τη δυνατότητα να *αντιμετωπίζουν επιτυχώς τα ορθογραφικά λάθη*. Σημείο αναφοράς στην κατηγορία αυτή αποτελεί η **μετρική edit distance**. Η μετρική αυτή υπολογίζει τη διαφορά δυο συμβολοσειρών, σ_1 και σ_2, με βάση τον αριθμό των επιτρεπόμενων μετασχηματισμών μεμονωμένων χαρακτήρων (εισαγωγή, διαγραφή ή αντικατάσταση χαρακτήρα) που απαιτούνται για να ταυτιστούν οι σ_1 και σ_2. Αποδίδοντας διαφορετικό κόστος σε κάθε μια από τις πράξεις αυτές, προκύπτει μια ποικιλία παρεμφερών μετρικών, με απλούστερη τη **Levenstein distance** που θεωρεί όλες τις πράξεις ισοδύναμες (με κόστος 1). Καλά αποτελέσματα δίνουν επίσης η **μετρική Jaro**, η οποία έχει κυρίως εφαρμογή στη σύγκριση ονομάτων και επωνύμων, καθώς επίσης και μια παραλλαγή της που αποδίδει μεγαλύτερο βάρος στην ταύτιση των προθεμάτων των λέξεων, η **μετρική Jaro-Winkler**.

- **αναπαράσταση με σύνολο λέξεων** (*tokens*)

  Οι μετρικές της κατηγορίας αυτής αντιμετωπίζουν τις συμβολοσειρές ως ένα σύνολο από λέξεις. Για το λόγο αυτό επιτυγχάνουν *καλύτερη απόδοση όταν οι συγκρινόμενες συμβολοσειρές αποτελούνται από δύο ή περισσότερες λέξεις καθώς είναι σε θέση να εντοπίζουν/αγνοούν τυχόν μεταθέσεις των λέξεων τους*. Η απλούστερη μετρική της κατηγορίας αυτής είναι η **ομοιότητα Jaccard**. Αυτή ορίζεται ως ο λόγος του πλήθους των κοινών λέξεων προς το πλήθος όλων των λέξεων (οι κοινές λέξεις υπολογίζονται μια φορά). Άλλες αξιόλογες μετρικές είναι η **Fellegi-Sunter** και η **Jensen-Shannon**, ενώ την καλύτερη επίδοση παρουσιάζει σε όλα τα πειράματα η **TFIDF**. Επειδή οι ορισμοί των μετρικών αυτών ξεφεύγουν από τους σκοπούς της εργασίας, παραπέμπουμε τον αναγνώστη στην εργασία [CRF03].

- **υβριδικές μέθοδοι**

  Οι μέθοδοι αυτές επιχειρούν να συνδυάσουν τα πλεονεκτήματα των δυο προηγούμενων κατηγοριών για να επιτύχουν ακόμη καλύτερη απόδοση. Πράγματι, η **SoftTFIDF** που ανήκει στην κατηγορία αυτή εμφανίζεται να έχει την καλύτερη απόδοση από όλες τις μετρικές ανεξαρτήτως κατηγορίας. Η μετρική αυτή προτείνεται στο [CRF03] ως συνδυασμός της TFIDF με κάποια από τις μετρικές της πρώτης κατηγορίας ώστε να βελτιώσει την απόδοσή της σε περιπτώσεις τυπογραφικών λαθών. Στην ίδια

δημοσίευση προτείνεται και μια άλλη ομάδα υβριδικών μεθόδων βασισμένη σε μια πρόταση των Monge-Elkan, με φτωχότερα όμως αποτελέσματα.

- **Q-grams**

  Η μέθοδος αυτή δεν κατατάσσεται σε καμία από τις προαναφερθείσες κατηγορίες, χρήζει όμως αναφοράς λόγω της ιδιαίτερα καλής απόδοσής της και της ευρείας εφαρμογής της στις βάσεις δεδομένων. Για να εξετάσει το βαθμό ταιριάσματος δυο συμβολοσειρών, διολισθαίνει ένα *παράθυρο μήκους q* πάνω από τους χαρακτήρες μιας συμβολοσειράς και μετρά τον αριθμό των q-grams που ταιριάζουν στη δεύτερη συμβολοσειρά. Περισσότερα για τη μέθοδο αυτή μπορεί να βρει ο ενδιαφερόμενος αναγνώστης στην εργασία [Ραπ07].

Προσαρμόζοντας τα πλεονεκτήματα των παραπάνω κατηγοριών στην περίπτωσή μας, θα λέγαμε ότι η μεν πρώτη αποτελεί την καλύτερη λύση για σύγκριση ονομάτων, ενώ η δεύτερη και η τρίτη είναι ιδανικές για σύγκριση τίτλων δημοσιεύσεων και τίτλων περιοδικών/συνδερίων. Το σημαντικότερο όμως πλεονέκτημα των μεθόδων αυτών (εκτός φυσικά από την εξαιρετική ακρίβεια που επιτυγχάνει η κάθε μια στον τομέα που εξειδικεύεται) είναι το γεγονός ότι δεν χρειάζονται καμία εκπαίδευση. Δεν απαιτείται δηλαδή καμία εκ των προτέρων γνώση για τις συγκρινόμενες συμβολοσειρές για να επιτύχουμε την επιθυμητή ακρίβεια.

Στο κεφάλαιο 3 αναφέρουμε τις μεθόδους που επιλέξαμε να χρησιμοποιήσουμε στην προσέγγισή μας.

2.2.5 Εξαγωγή πληροφορίας (Information Extraction) & Wrapper Maintenance

Η υποενότητα αυτή είναι αφιερωμένη σε κάποια θέματα που σχετίζονται με τις ακαδημαϊκές μηχανές αναζήτησης και στα οποία εφαρμόζονται μέθοδοι μηχανικής μάθησης. Εξαιρετικές πηγές πληροφοριών για τα θέματα αυτά αποτελούν οι εργασίες [Συγ05] και [Τσου06].

Αντικείμενο της **εξαγωγής πληροφορίας** είναι η ανάπτυξη συστημάτων τα οποία με βάση ένα σύνολο κανόνων εξάγουν δεδομένα από έγγραφα διαφόρων τύπων. Πιο συγκεκριμένα, διακρίνουμε τους εξής τύπους εγγράφων:

- *Μη δομημένα*, όπως ελεύθερο κείμενο
- *Δομημένα*, για παράδειγμα XML αρχεία, βάσεις δεδομένων.
- *Ημιδομημένα*, όπως ιστοσελίδες γραμμένες σε γλώσσα HTML

Στα πλαίσια αυτής της εργασίας, μας ενδιαφέρουν συστήματα που εξάγουν δεδομένα από ημιδομημένα κείμενα, δηλαδή από ιστοσελίδες σε HTML, και τα μετατρέπουν σε δομημένη μορφή. Τα συστήματα αυτά ονομάζονται **wrappers**. Η λειτουργία τους βασίζεται σε ένα σύνολο κανόνων που εκμεταλλεύεται το layout των ιστοσελίδων και την κανονικότητα που

αυτό εμφανίζει ως προς τα html tags, για να αποσπάσει την πληροφορία που αυτά εμπερικλείουν.

Υπάρχουν δυο τρόποι ανάπτυξης ενός wrapper:

- *Χειρωνακτικά*, με τη βοήθεια κάποιας γλώσσας προγραμματισμού (π.χ. JAVA) χρησιμοποιώντας κανονικές εκφράσεις. Η μέθοδος αυτή στιγματίζεται από τη δυσκολία και τις μεγάλες απαιτήσεις της σε χρόνο, κυρίως όταν πρόκειται για μεγάλο αριθμό ιστοσελίδων.
- *Αυτόματα ή ημιαυτόματα*, με τη βοήθεια κατάλληλων εργαλείων, μια καταγραφή των οποίων επιχειρείται στο [LRST02]. Σπάνια ωστόσο τα αποτελέσματα των εργαλείων αυτών είναι ικανοποιητικά, ενώ φαίνεται ότι και η σχετική έρευνα βρίσκεται σε τέλμα.

Εξαιρετικά δύσκολη καθίσταται ωστόσο και η συντήρηση ενός wrapper, το γνωστό και ως **Wrapper Maintenance Problem**, λόγω των συχνών αλλαγών στη δομή των ιστοσελίδων που επεξεργάζεται. Αυτό έχει ως αποτέλεσμα να εξάγονται εσφαλμένα δεδομένα σε περίπτωση αλλαγών. Η επίλυσή του προβλήματος χωρίζεται σε δύο σκέλη:

- **Wrapper Verification**, δηλαδή την επαλήθευση της ορθής λειτουργίας του wrapper. Έτσι καθίσταται δυνατός ο εντοπισμός αλλαγών στη δομή της ιστοσελίδας.
- **Wrapper Reinduction**, δηλαδή την επαγωγή ενός νέου σωστού wrapper σε περίπτωση που εντοπιστεί αλλαγή.

Η προσέγγισή μας στα παραπάνω ζητήματα (ανάπτυξη και συντήρηση wrapper) αναλύεται στο κεφάλαιο 4, παράγραφος 4.3.4.

3

Ανάκτηση δεδομένων από το Web
Επίλυση των προβλημάτων CMP, MCP & SCP

Το κεφάλαιο αυτό είναι αφιερωμένο στο σημαντικότερο κομμάτι αυτής της εργασίας, δηλαδή την ανάκτηση και εξαγωγή πληροφορίας από ακαδημαϊκές μηχανές αναζήτησης (το Google Scholar εν προκειμένω). Η ιδιαίτερη σημασία της διαδικασίας αυτής οφείλεται στο γεγονός ότι αποτελεί το μοναδικό (προς το παρόν) τρόπο εισαγωγής πληροφοριών στη βάση δεδομένων. Παράλληλα με την περιγραφή μιας αναζήτησης δεδομένων στο Web θα παρουσιάσουμε και θα αναλύσουμε λεπτομερώς τους αλγορίθμους μηχανικής μάθησης που αναπτύξαμε για την επίλυση των προβλημάτων citation matching, mixed και split citation. Τους αλγορίθμους δηλαδή που εξασφαλίζουν την ορθότητα των δεδομένων που παρουσιάζονται στο χρήστη και καταχωρούνται στη βάση δεδομένων.

Κατ' αρχήν, πρέπει να τονίσουμε ότι κάθε on-line αναζήτηση αποτελείται από τα ακόλουθα διαδοχικά στάδια:

- Εύρεση των δημοσιεύσεων και των ενδεχόμενων συνώνυμών που αντιστοιχούν στο δοσμένο όνομα επιστήμονα (s*plit citation problem*)
- Ανάκτηση των δημοσιεύσεων που αντιστοιχούν σε κάθε ένα από τα *επιβεβαιωμένα* συνώνυμα του συγκεκριμένου επιστήμονα
- Εντοπισμός των δημοσιεύσεων που στην πραγματικότητα αναφέρονται στο ίδιο paper ή βρίσκονται ήδη καταχωρημένες στη βάση δεδομένων (*citation matching problem*)
- Επεξεργασία των δημοσιεύσεων από το χρήστη και καταχώρησή τους στη βάση δεδομένων
- Αναγνώριση των διαφορετικών επιστημόνων που συμμετέχουν στις καταχωρημένες δημοσιεύσεις (*mixed & split citation problem*)
- Επεξεργασία των διαφορετικών επιστημόνων από το χρήστη και καταχώρησή τους στη βάση δεδομένων
- Αναζήτηση των βιβλιογραφικών αναφορών κάθε δημοσίευσης του δοσμένου επιστήμονα

Οι επιμέρους αυτές διεργασίες θα αποσαφηνισθούν στις επόμενες ενότητες, κάθε μια από τις οποίες είναι αφιερωμένη σε μια διεργασία.

3.1 Εύρεση δημοσιεύσεων και πιθανών συνωνύμων για το δοσμένο όνομα επιστήμονα

Το πρώτο βήμα μιας on-line αναζήτησης συνίσταται στην *ανάκτηση όλων των δημοσιεύσεων που αντιστοιχούν στο δοσμένο όνομα επιστήμονα*. Τη διαδικασία αυτή αναλαμβάνει ο wrapper του Google Scholar. Όπως αναφέρθηκε και προηγουμένως, ο wrapper αυτός ανακτά διαδοχικά τις HTML σελίδες των αποτελεσμάτων και τις διασπά σε εγγραφές, κάθε μια από τις οποίες αντιστοιχεί σε μια δημοσίευση. Στη συνέχεια, ανάλογα με την κατηγορία στην οποία κατατάσσεται η εγγραφή, συμπληρώνει τα πεδία ενός αντικειμένου της κλάσης Paper με τις αντίστοιχες πληροφορίες.

Το επόμενο βήμα αποτελεί μια πρωτότυπη προσπάθεια επίλυσης του split citation problem, μόνο όμως όσον αφορά το δοσμένο όνομα επιστήμονα. Πιο συγκεκριμένα, πρόκειται για μια προσπάθεια *εντοπισμού πιθανών συνωνύμων* βασισμένη στο γεγονός ότι *μια αναζήτηση με τον τίτλο μιας δημοσίευσης μπορεί να μας αποφέρει συνώνυμα για τους συγγραφείς της*. Αυτό το φαινόμενο αποδίδεται κυρίως στα παρακάτω αίτια :

- Κάνοντας μια αναζήτηση στο Google Scholar για τις δημοσιεύσεις που περιέχουν στο τίτλο τους συγκεκριμένες λέξεις, μπορεί να βρει κανείς παραλλαγές μιας δημοσίευσης που δε εμφανίζονται σε μια αναζήτηση με βάση το όνομα ενός από τους συγγραφείς της. Αυτό οφείλεται συνήθως στο ότι οι τίτλοι τέτοιων παραλλαγών διαφέρουν σε σημαντικό βαθμό από τον κανονικό. Διαφορετικά το Google Scholar θα είχε εντοπίσει την ταύτιση και η συγκεκριμένη δημοσίευση θα συμπεριλαμβανόταν με τα αποτελέσματα της αρχικής αναζήτησης. Σε τέτοιες περιπτώσεις είναι εξαιρετικά πιθανό να διαφέρουν και τα ονόματα των συγγραφέων, π.χ. εξαιτίας λανθασμένης επεξεργασίας-parsing από το Google Scholar. Ένα τυπικό παράδειγμα τέτοιων δημοσιεύσεων έχουμε στις περιπτώσεις όπου τα ονόματα ενός ή περισσότερων συγγραφέων προσκολλώνται στην αρχή ή στο τέλος του τίτλου μιας δημοσίευσης (βλέπε εικόνα 2).
- Εξίσου συχνό είναι το φαινόμενο κατά το οποίο δυο ή περισσότερες βιβλιογραφικές αναφορές έχουν ακριβώς τον ίδιο τίτλο αλλά να αναφέρουν τους συγγραφείς με διαφορετικά ονόματα. Το Google Scholar αντιστοιχώντας τις συγκεκριμένες βιβλιογραφικές αναφορές στην ίδια δημοσίευση, επιλέγει αναγκαστικά ένα όνομα για κάθε συγγραφέα, αγνοώντας τις παραλλαγές των ονομάτων τους. Η αναζήτηση όμως

με βάση τον τίτλο της δημοσίευσης επιστρέφει ως διακεκριμένες αυτές τις βιβλιογραφικές αναφορές και με αυτόν τον τρόπο είμαστε σε θέση να εντοπίσουμε τα συνώνυμα των συγγραφέων.

Στη συνέχεια παραθέτουμε ένα ενδεικτικό παράδειγμα για του λόγου το αληθές:


An **evaluation** of **Naive Bayesian anti-spam filtering** - all 15 versions »
... , J Koutsias, KV Chandrinos, G **Paliouras**, CD ... - Arxiv preprint cs.CL/0006013, 2000 - arxiv.org
... Learning, Barcelona, Spain, pp. 9-17, 2000. An **Evaluation** of **Naive Bayesian Anti-Spam Filtering** Ion Androutsopoulos, John Koutsias **...**
Cited by 152 - Related Articles - View as HTML - Web Search - Import into BibTeX - Library Search

[CITATION] G. and Spyropoulos, CD 2000. An **Evaluation** of **Naive Bayesian Anti-Spam Filtering**
IK Androutsopoulos, J Chandrinos, KV **Paliouras** - Proc. of the workshop on Machine Learning in the New ...
Cited by 2 - Related Articles - Web Search - Import into BibTeX

Εικόνα 2. Παραλλαγές μιας δημοσίευσης στο Google Scholar

Στην περίπτωση αυτή, το πραγματικό όνομα, G Paliouras, έμφανίζεται στη δεύτερη περίπτωση ως KV Paliouras, καθώς κατά την επεξεργασία της δεύτερης δημοσίευσης, ο αλγόριθμος του Google Scholar αντιστοίχησε σε κάθε επώνυμο το όνομα του προηγούμενού συγγραφέα.

Ο αλγόριθμος που αναπτύξαμε για να εκμεταλλευτούμε τέτοια φαινόμενα έχει ως εξής:

Για κάθε μια από τις δημοσιεύσεις που βρέθηκαν με το αρχικά δοσμένο όνομα, εκτελούμε μια νέα αναζήτηση στο Google Scholar. Κατά την αναζήτηση αυτή βρίσκουμε τις δημοσιεύσεις που έχουν στον τίτλο τους *όλες* τις λέξεις της τρέχουσας δημοσίευσης.

Για κάθε μια από τις δημοσιεύσεις των αποτελεσμάτων, βρίσκουμε μόνο τους συγγραφείς της και στη συνέχεια εξετάζουμε αν ταιριάζουν με τους συγγραφείς της τρέχουσας δημοσίευσης.

Αν η ομοιότητα υπερβαίνει ένα κατώφλι που έχουμε θέσει, επιλέγουμε τον συγγραφέα εκείνο που ταιριάζει περισσότερο με το αρχικά δοσμένο όνομα συγγραφέα. Αν η αντίστοιχη μέγιστη ομοιότητα ξεπερνάει και αυτή το σχετικό κατώφλι, θεωρούμε το όνομα αυτό συνώνυμο του συγγραφέα. Εξασφαλίζουμε, βέβαια, πρώτα ότι το νέο αυτό συνώνυμο δεν υπάρχει ήδη στο σύνολο των συνωνύμων, καθώς σε αυτό διατηρούμε μόνο διαφορετικά ονόματα.

Για τον υπολογισμό ομοιότητας στις δυο προαναφερθείσες περιπτώσεις, χρησιμοποιούμε τις εξής μετρικές ταιριάσματος συμβολοσειρών :

- *SoftTFIDF σε συνδυασμό με τη Jaro με σχετικά χαμηλό κατώφλι* (0.7 και για τις δυο) για τη σύγκριση των συγγραφέων δυο δημοσιεύσεων. Πιο αναλυτικά, για τις ανάγκες της συγκεκριμένης σύγκρισης, τοποθετούμε τα ονόματα των συγγραφέων κάθε δημοσίευσης σε μια συμβολοσειρά. Οι δυο συμβολοσειρές που προκύπτουν θεωρούμε ότι αποτελούν στην ουσία σύνολα λέξεων. Συνεπώς, για τον υπολογισμό της ομοιότητάς τους πρέπει να επιλέξουμε μια μετρική που ειδικεύεται στη σύγκριση tokens. Όπως προκύπτει τόσο από το [CRF03] όσο και από το [Ραπ07], οι καλύτερες μετρικές για τέτοιου είδους συγκρίσεις είναι η TFIDF και η παραλλαγή της SoftTFIDF.

Επειδή όμως θέλουμε να επιτρέψουμε σημαντικές διαφοροποιήσεις στα ονόματα των δυο συμβολοσειρών (π.χ. ορθογραφικά λάθη) επιλέξαμε τη δεύτερη, την οποία συνδυάσαμε με τη μετρική Jaro. Έτσι, σε αντίθεση με την TFIDF που απαιτεί απόλυτη ταύτιση για τα tokens, το τελικό αποτέλεσμα δεν επηρεάζεται από τις παραλλαγές των ονομάτων των συγγραφέων. Άλλωστε, στόχος μας είναι η εύρεση συνωνύμων. Γι' αυτό και θέσαμε το κατώφλι της Jaro στο 0.7 ώστε να μην αγνοούνται σημαντικές διαφοροποιήσεις των ονομάτων, μέσα σε λογικά πλαίσια πάντα. Ομοίως, χαμηλά θέσαμε και το κατώφλι της SoftTFIDF, ώστε να μην απορρίπτονται περιπτώσεις που ένας συγγραφέας έχει αφαιρεθεί ή ένα άλλος έχει προστεθεί. Συνολικά, η μετρική αυτή ουσιαστικά επιστρέφει *το ποσοστό των συγγραφέων που εμφανίζονται και στις δυο δημοσιεύσεις*, έστω και με ελαφρώς διαφορετικά ονόματα.

- *Jaro με κατώφλι 0.7* για τον εντοπισμό του πιο όμοιου με τον αρχικά δοσμένο συγγραφέα. Αυτή η επιπλέον σύγκριση είναι απαραίτητη αν αναλογιστεί κανείς ότι δεν αρκεί η ομοιότητα μεταξύ των συγγραφέων δυο δημοσιεύσεων να ικανοποιεί το πρώτο κριτήριο για να θεωρηθεί ότι βρήκαμε συνώνυμο. Αντίθετα, πρέπει και το πιο κοντινό στο αρχικά δοσμένο όνομα να είναι αρκετά όμοιο με αυτό για να θεωρηθεί συνώνυμο. Και στη συγκεκριμένη περίπτωση όμως, το κατώφλι είναι σχετικά χαμηλό για να επιτρέπει σημαντικές, σε λογικά όμως πλαίσια, αποκλίσεις μεταξύ των συνωνύμων.

3.2 Ανάκτηση των δημοσιεύσεων που αντιστοιχούν σε κάθε επιβεβαιωμένο συνώνυμο

Στο στάδιο αυτό αρχικά ζητείται από το χρήστη να *επιβεβαιώσει τα αποτελέσματα της προηγούμενης διαδικασίας*. Παρουσιάζεται δηλαδή στο χρήστη ένα παράθυρο στο οποίο εμφανίζεται το σύνολο των διαφορετικών ονομάτων-συνωνύμων που εντοπίσθηκαν από τον παραπάνω αλγόριθμο. Αν βέβαια δεν έχει βρεθεί κανένα συνώνυμο, αυτή η διαδικασία παρακάμπτεται και μεταβαίνουμε στο επόμενο στάδιο. Στην αντίθετη περίπτωση, ο χρήστης καλείται να επιλέξει όσα από τα συνώνυμα θεωρεί ότι πράγματι αφορούν τον αρχικά δοσμένο επιστήμονα, ώστε να αγνοηθούν τα υπόλοιπα και να αποφευχθούν περιττές αναζητήσεις. Το αποτέλεσμα δηλαδή της προηγούμενης διεργασίας υπόκειται στον τελικό έλεγχο του χρήστη. Άλλωστε, λόγω τον σχετικά ελαστικών κατωφλιών στις χρησιμοποιούμενες μετρικές, είναι πιθανόν ορισμένα ονόματα να θεωρηθούν λανθασμένα συνώνυμα ακόμα και αν είναι προφανές (για έναν άνθρωπο) ότι δε σχετίζονται σε καμία περίπτωση με το δοσμένο επιστήμονα. Το φαινόμενο αυτό παρατηρείται, για παράδειγμα, κατά τη σύγκριση ελληνικών επιθέτων με κοινή κατάληξη, τα οποία τυχαίνει επίσης να ταιριάζουν σε κάποια άλλα

γράμματα εκτός της κατάληξης. Συνήθως, όμως, οι σχετικές μετρικές, της κατηγορίας "αναπαράσταση με χαρακτήρες" (ενότητα 2.2.4), αποδίδουν υψηλή ομοιότητα σε τέτοιες συγκρίσεις, αδυνατώντας να εντοπίσουν τέτοιες προφανείς αναντιστοιχίες. Ενδεικτικά αναφέρουμε ότι η τιμή ομοιότητας που επιστρέφει η μετρική Jaro για τα ονόματα P Stamatopoulos και M Hatzopoulos είναι 0.83, αισθητά πάνω από το κατώφλι που έχουμε θέσει.

Παρ' όλα αυτά, πιστεύουμε ότι *η τελική επιβεβαίωση από το χρήστη, δεν μειώνει τη χρηστικότητα της εφαρμογής*. Και αυτό γιατί αφενός τέτοιες περιπτώσεις είναι σπάνιες και αφετέρου στην κρίση του χρήστη αφήνεται συνήθως ένας μικρός αριθμός συνωνύμων. Μόνο δηλαδή ένα πολύ μικρό μέρος των συγκρίσεων χρειάζεται να επιβεβαιωθεί από το χρήστη, αφού η συντριπτική τους πλειοψηφία έχει διεκπεραιωθεί αυτόματα από τον αντίστοιχο αλγόριθμο.

Μετά την επιλογή των ορθών συνωνύμων από το χρήστη, ξεκινάει *η διαδικασία εύρεσης των δημοσιεύσεων που αντιστοιχούν σε καθένα από αυτά*. Πραγματοποιούνται δηλαδή νέες αναζητήσεις και οι βιβλιογραφικές αναφορές που προκύπτουν από αυτές προστίθενται στις αρχικές. Απαραίτητη προϋπόθεση ωστόσο για να συμπεριληφθούν στις αρχικές δημοσιεύσεις, είναι να έχουν διαφορετικό τίτλο από τις ήδη υπάρχουσες. Όσες έχουν τίτλο ίδιο με κάποια άλλη δημοσίευση απορρίπτονται με το σκεπτικό ότι κατά πάσα πιθανότητα ταυτίζονται, λόγω της επεξεργασίας που κάνει το Google Scholar. Είναι δηλαδή πρακτικά αδύνατο δυο εργασίες να έχουν ακριβώς τον ίδιο τίτλο και παρόμοιους συγγραφείς και ο αλγόριθμος του Google Scholar να τις θεωρεί διαφορετικές. Πρέπει να σημειώσουμε στο σημείο αυτό ότι πριν τις απορρίψουμε, καλούμε μια συνάρτηση "συγχώνευσης" των ταυτόσημων δημοσιεύσεων και κρατάμε την δημοσίευση που προκύπτει από αυτή τη διαδικασία. Με το όρο "συγχώνευση" εννοούμε τον συνδυασμό των πληροφοριών των δυο βιβλιογραφικών αναφορών, ώστε να διατηρηθούν τυχόν νέες πληροφορίες της υπό απόρριψη δημοσίευσης, όπως ένα επιπλέον URL. Τέλος, πρέπει να επισημάνουμε ότι η διαδικασία εύρεσης συνωνύμων δεν επαναλαμβάνεται για τις νέες δημοσιεύσεις. Εκτελείται δηλαδή μόνο μια φορά, αποκλειστικά για τις δημοσιεύσεις που αντιστοιχούν στο αρχικά δοσμένο όνομα επιστήμονα.

3.3 Επίλυση του Citation Matching προβλήματος

Όπως αναφέραμε και στην σχετική υποενότητα του 2ου κεφαλαίου (2.2.3), στόχος του συγκεκριμένου σταδίου είναι *ο εντοπισμός και η εξάλειψη των διπλότυπων δημοσιεύσεων*.

Των δημοσιεύσεων δηλαδή που εμφανίζονται ως διαφορετικές αλλά στην πραγματικότητα αντιστοιχούν στην ίδια δημοσίευση. Όπως θα δούμε και στη συνέχεια, μερική ή ελλιπής επίλυση του προβλήματος αυτού έχει σημαντική επίπτωση στην ποιότητα των πληροφοριών που καταχωρούνται στη βάση δεδομένων και παρουσιάζονται στο χρήστη.

Θεωρώντας ότι το πρόβλημα αυτό έχει επιλυθεί για τις δημοσιεύσεις που είναι καταχωρημένες στη ΒΔ, περιορίζουμε το πεδίο εφαρμογής της παρακάτω μεθόδου στις νέες δημοσιεύσεις που προέκυψαν από την τελευταία αναζήτηση, συμπεριλαμβανομένων και όσων προέκυψαν από την αναζήτηση με βάση τα επιβεβαιωμένα συνώνυμα.

Ο αλγόριθμος που αναπτύξαμε για τη συγκεκριμένη διεργασία λειτουργεί ως εξής :

Κατ' αρχήν *ταξινομούμε τις δημοσιεύσεις σε φθίνουσα σειρά ανάλογα με το πλήθος των citations τους*. Αυτό γίνεται με το σκεπτικό ότι όσα περισσότερα citations έχει μια δημοσίευση, τόσο μεγαλύτερη είναι η πιθανότητα να είναι σωστά τα στοιχεία της και δη ο τίτλος της. Αυτό βέβαια ισχύει κυρίως για παραλλαγές της ίδιας δημοσίευσης (η παραλλαγή με τα περισσότερα citations είναι σχεδόν σίγουρο ότι έχει τα πιο σωστά στοιχεία) και όχι μεταξύ διαφορετικών δημοσιεύσεων. Δε σημαίνει δηλαδή ότι μια δημοσίευση με 1 ή κανένα citation έχει πάντα λανθασμένα δεδομένα ή αντίθετα ότι μια δημοσίευση με 100 ή περισσότερα citations έχει οπωσδήποτε σωστά στοιχεία. *Στην περίπτωση που δυο δημοσιεύσεις έχουν τον ίδιο αριθμό citations, η ταξινόμησή τους γίνεται με βάση το πλήθος των URL πάλι σε φθίνουσα σειρά*. Η αιτιολογία και για το κριτήριο αυτό, είναι ίδια με την προηγούμενη. Το γεγονός δηλαδή ότι το Google Scholar έχει βρει μια δημοσίευση διαθέσιμη σε περισσότερους δικτυακούς τόπους απ' ότι μια παραλλαγή είναι ισχυρή ένδειξη ότι η πρώτη δημοσίευση έχει περισσότερες πιθανότητες να έχει σωστά δεδομένα.

Ταξινομώντας επομένως τις δημοσιεύσεις με αυτόν τον τρόπο και *συγκρίνοντας κάθε δημοσίευση με την προηγούμενή της*, εξασφαλίζουμε ότι όσο μεγαλύτερη είναι η πιθανότητα ορθότητας μιας βιβλιογραφικής αναφοράς, με τόσο λιγότερες δημοσιεύσεις συγκρίνεται. Αντίθετα, οι δημοσιεύσεις με μεγάλη πιθανότητα λανθασμένων δεδομένων συγκρίνονται με το μεγαλύτερο αριθμό δημοσιεύσεων. Διευκολύνεται επίσης και ο τελικός συνδυασμός-συγχώνευση των ταυτόσημων βιβλιογραφικών αναφορών, καθώς γνωρίζουμε ότι σωστότερα δεδομένα έχει εκείνη που εμφανίζεται πιο ψηλά στην ταξινομημένη σειρά.

Για κάθε μια από τις ταξινομημένες δημοσιεύσεις κάνουμε διαδοχικά τους εξής ελέγχους:

- Αρχικά *εξετάζουμε αν είναι καταχωρημένη στη βάση δεδομένων*, γιατί στην περίπτωση αυτή έχει ήδη επιλυθεί από προηγούμενη αναζήτηση το citation matching problem. Δε χρειάζεται, επομένως, να επαναλαμβάνεται άσκοπα η ίδια διαδικασία.

  Ο έλεγχος αυτός υλοποιείται με τα εξής βήματα :

  Ανακαλούμε από τη βάση δεδομένων *όλες* τις δημοσιεύσεις που έχουν τον ίδιο ακριβώς τίτλο ή που τους έχει αντιστοιχηθεί ο συγκεκριμένος τίτλος ως παραλλαγή κατά το citation matching.

Στη συνέχεια, εάν έχουν βρεθεί περισσότερες από μια δημοσιεύσεις, συγκρίνουμε τους συγγραφείς κάθε μιας με αυτούς της τρέχουσας δημοσίευσης. Η σύγκριση αυτή γίνεται με μια μετρική SoftTFIDF που χρησιμοποιεί τη Jaro με κατώφλι 0.8, όσο είναι δηλαδή και το κατώφλι για την εύρεση συνωνύμων κατά την επίλυση των προβλημάτων mixed και split citation (βλέπε παρακάτω).

Αν η ομοιότητα των συγγραφέων υπερβαίνει ένα αρκετά χαμηλό κατώφλι (0.3), θεωρούμε ότι οι δύο δημοσιεύσεις ταυτίζονται. Στο σημείο αυτό η επεξεργασία για τη συγκεκριμένη δημοσίευση σταματάει και, επιπλέον, δεν συμπεριλαμβάνουμε τη συγκεκριμένη ταύτιση σε αυτές που ζητάμε από το χρήστη να επιβεβαιώσει.

Ένας από τους λόγους για το παραπάνω, πολύ χαμηλό όριο επιβεβαίωσης της ταύτισης των δημοσιεύσεων είναι το γεγονός ότι πολύ σπάνια οι τίτλοι δυο διαφορετικών δημοσιεύσεων ταυτίζονται απόλυτα. Θα χρειαστεί δηλαδή σε πολύ λίγες περιπτώσεις να γίνει αυτή η σύγκριση των συγγραφέων τους. Επιπλέον όμως οι παραλλαγές του τίτλου μιας δημοσίευσης συνδέονται, μετά από την καταχώρηση τους στη βάση δεδομένων, με τους συγγραφείς που αντιστοιχούν στο σωστό τίτλο και όχι με τους δικούς τους συγγραφείς. Συνεπώς, κατά τη σύγκριση των συγγραφέων η ομοιότητα είναι εξαιρετικά χαμηλή, ιδιαίτερα όταν πρόκειται για παραλλαγές με λανθασμένους ή ελλιπείς συγγραφείς. Η κατάσταση αυτή θα διευκρινιστεί εξετάζοντας ως παράδειγμα την ακόλουθη εικόνα.


An **evaluation** of **Naive Bayesian anti-spam filtering** - all 15 versions »
... , J Koutsias, KV Chandrinos, G **Paliouras**, CD ... - Arxiv preprint cs.CL/0006013, 2000 - arxiv.org
... Learning, Barcelona, Spain, pp. 9-17, 2000. An **Evaluation** of **Naive Bayesian Anti-Spam Filtering** Ion Androutsopoulos, John Koutsias **...**
Cited by 152 - Related Articles - View as HTML - Web Search - Import into BibTeX - Library Search

[CITATION] G. and Spyropoulos, CD 2000. An **Evaluation** of **Naive Bayesian Anti-Spam Filtering**
IK Androutsopoulos, J Chandrinos, KV **Paliouras** - Proc. of the workshop on Machine Learning in the New ...
Cited by 2 - Related Articles - Web Search - Import into BibTeX

Εικόνα 3. Παραλλαγές μιας δημοσίευσης στο Google Scholar

Στην περίπτωση αυτή, η δεύτερη δημοσίευση έχει και λανθασμένους και ελλιπείς συγγραφείς, καθώς ένας από αυτούς έχει προστεθεί στον τίτλο της. Κατά την καταχώρησή της όμως στη βάση δεδομένων θα συσχετιστεί με τους (σωστούς) συγγραφείς της πρώτης δημοσίευσης. Έτσι, σε επόμενες αναζητήσεις, αφού εντοπιστεί ότι ο τίτλος της δεύτερης δημοσίευσης είναι καταχωρημένος στη βάση δεδομένων, θα πρέπει συγκριθούν οι συγγραφείς της με αυτούς της πρώτης για να επιβεβαιωθεί η ταύτιση. Η ομοιότητα όμως που θα αποδώσει η σχετική μετρική είναι εξαιρετικά χαμηλή.

- Αν η τρέχουσα δημοσίευση δεν υπάρχει ήδη καταχωρημένη στη βάση δεδομένων, τη *συγκρίνουμε με **όλες** τις προηγούμενές στην κατάταξη* νέες βιβλιογραφικές αναφορές, όπως αυτές προέκυψαν από την προηγούμενη ταξινόμηση. Ο έλεγχος αυτός γίνεται με

βάση την ομοιότητα των τίτλων, χρησιμοποιώντας πάλι μια εκδοχή της SoftTFIDF που χρησιμοποιεί τη μετρική Jaro και κατώφλι 0.8. Όπως και στην προηγούμενη σύγκριση, εντοπίζουμε τον τίτλο με τη μεγαλύτερη ταύτιση και στη συνέχεια εξετάζουμε αν η ομοιότητά τους υπερβαίνει ένα κατώφλι, 0.7 στη συγκεκριμένη περίπτωση. Το κατώφλι είναι σχετικά ελαστικό γιατί σκοπός μας είναι να μη διαφεύγουν από τη σύγκριση αυτή οι πιο συνηθισμένες περιπτώσεις παραλλαγής του τίτλου. Αυτές είναι τα ορθογραφικά λάθη και η προσθήκη συνεδρίου, χρονολογίας ή συγγραφέα στην αρχή ή στο τέλος του τίτλου της δημοσίευσης. Άλλωστε, πρέπει να σημειώσουμε ότι τα ταιριάσματα που προκύπτουν από τον έλεγχο αυτό ελέγχονται και επικυρώνονται από το χρήστη στο επόμενο στάδιο. Έτσι, τυχόν προφανή λάθη αποφεύγονται εύκολα. Στη συνέχεια παραθέτουμε ενδεικτικά παραδείγματα για τις **πιο συνηθισμένες περιπτώσεις παραλλαγής του τίτλου μιας δημοσίευσης** (τα παραδείγματα προκύπτουν από την επεξεργασία των δημοσιεύσεων του P Faloutsos) :

- ορθογραφικό λάθος

  On power law relations of the Internet Topology
  On power-law relationships of the Internet topology
  SoftTFIDF score = 0.987179487179487

  BOn PowerYLaw Relationships of the Internet Topology
  On power-law relationships of the Internet topology
  SoftTFIDF score = 0.9007693708900227

- προσθήκη ονομάτων συγγραφέων στην αρχή του τίτλου

  Van de panne M, Terzopoulos D. Composable controllers for physics-based character animation
  Composable controllers for physics-based character animation
  SoftTFIDF score = 0.7337993857053426

- προσθήκη συνεδρίου στο τέλος του τίτλου

  On power-law relationships of the Internet topology, ACM SIG-COMM'99
  On power-law relationships of the Internet topology
  SoftTFIDF score = 0.8164965809277261

- Στην περίπτωση όπου ούτε από τον προηγούμενο έλεγχο προκύπτει κάποιο πιθανό ταίριασμα, *εξετάζουμε αν η τρέχουσα δημοσίευση ανήκει στην κατηγορία εκείνη των δημοσιεύσεων που ο τίτλος τους έχει τοποθετηθεί σε διαφορετικό πεδίο* (είτε ως όνομα συγγραφέα είτε ως τίτλος περιοδικού). Τα κριτήρια που συνηγορούν σε αυτό το

ενδεχόμενο είναι το *πλήθος των URL* σε συνδυασμό με τις *λέξεις του τίτλου ή του abstract* της υπό εξέταση βιβλιογραφικής αναφοράς.

Για την ακρίβεια, *αν για την τρέχουσα δημοσίευση έχουν βρεθεί λιγότερα από 2 URL και ο τίτλος ή το abstract της αποτελείται αντίστοιχα από λιγότερες από 3 ή 10 λέξεις*, τότε συγκρίνουμε ξανά την τρέχουσα δημοσίευση με όλες τις προηγούμενες με τον εξής όμως τρόπο : Αντί να συγκρίνουμε απλά τους τίτλους, εξετάζουμε την ομοιότητα των δημοσιεύσεων με βάση μια νέα συμβολοσειρά. Η συμβολοσειρά αυτή σχηματίζεται τοποθετώντας στην αρχή της τους συγγραφείς, έπειτα τον τίτλο και στο τέλος τον τίτλο του περιοδικού της εκάστοτε βιβλιογραφικής αναφοράς. Ωστόσο, η σειρά αυτή δεν επηρεάζει την ομοιότητα, καθώς οι μετρικές της αντίστοιχης κατηγορίας ("αναπαράσταση με σύνολο λέξεων") δε λαμβάνουν υπόψη τους τη σειρά των λέξεων.

Η μετρική που χρησιμοποιήσαμε και σε αυτήν την περίπτωση είναι η SoftTFIDF σε συνδυασμό με τη Jaro και κατώφλι 0.8. Για να θεωρηθεί ότι η πιο όμοια δημοσίευση που προκύπτει από αυτή τη σύγκριση πιθανώς ταυτίζεται με την τρέχουσα δημοσίευση, απαιτούμε η ομοιότητα τους να ξεπερνάει το 0.7.

Όσα ταιριάσματα εντοπιστούν με ένα από τους δυο τελευταίους τρόπους, επικυρώνονται στο επόμενο στάδιο από το χρήστη.

3.4 Επεξεργασία των δημοσιεύσεων από το χρήστη και καταχώρησή τους στη βάση δεδομένων

Το στάδιο αυτό ξεκινάει αμέσως μετά από την εύρεση των πιθανά ταυτόσημων βιβλιογραφικών αναφορών, οπότε και εμφανίζεται στο χρήστη το *παράθυρο επιβεβαίωσης*. Στο παράθυρο αυτό, εκτός από τα ζεύγη των δημοσιεύσεων που πρότεινε ο παραπάνω αλγόριθμος για "συγχώνευση", εμφανίζονται και όλες οι απαραίτητες πληροφορίες για να μπορέσει ο χρήστης να καταλήξει σε *ασφαλή συμπεράσματα* για την ορθότητα των προτάσεων αυτών. Πιο συγκεκριμένα, πατώντας πάνω στον τίτλο μιας δημοσίευσης εμφανίζεται ένα παράθυρο με όλα τα στοιχεία που εξήγαγε ο wrapper του Google Scholar για τις δημοσιεύσεις του συγκεκριμένου ζεύγους. Επιπλέον, παρέχονται στο χρήστη τα URLs τόσο για τις ιστοσελίδες όπου βρίσκεται διαθέσιμο το πλήρες κείμενο κάθε δημοσίευσης όσο και για τις ιστοσελίδες του Google Scholar που περιέχουν τα citations της. Κάθε φορά που επιλέγει ο χρήστης ένα από αυτά τα URLs, ανοίγει αυτόματα ένα παράθυρο ή tab του default browser στην επιλεγμένη ιστοσελίδα. Έτσι παρέχεται η δυνατότητα στο χρήστη να διαπιστώσει ποια δημοσίευση αφορά στην πραγματικότητα η συγκεκριμένη βιβλιογραφική αναφορά, πέρα από οποιοδήποτε λάθος στην επεξεργασία των βιβλιογραφικών αναφορών

που κάνει το Google Scholar. Αφού ο χρήστης επιλέξει ποια από τα προτεινόμενα ταιριάσματα είναι σωστά, οι αντίστοιχες δημοσιεύσεις "συγχωνεύονται".

Στο σημείο αυτό πρέπει να διευκρινίσουμε κάποια ζητήματα σχετικά με το παράθυρο επιβεβαίωσης των προτεινόμενων συγχωνεύσεων:

- Όπως διευκρινίσαμε και παραπάνω, *στο παράθυρο αυτό δεν εμφανίζονται δημοσιεύσεις που έχει διαπιστωθεί ότι είναι ήδη καταχωρημένες στη βάση δεδομένων*. Δηλαδή, η συγχώνευση μιας δημοσίευσης με μια ήδη υπάρχουσα γίνεται *αυτόματα*, χωρίς να ερωτηθεί ο χρήστης. Άλλωστε θεωρούμε αρκετά ασφαλή και εύρωστα τα κριτήρια που έχουμε θέσει για την περίπτωση αυτή.
- *Ενδέχεται στο παράθυρο αυτό να εμφανίζονται δυο δημοσιεύσεις που ταιριάζουν στον τίτλο, χωρίς αυτός όμως να είναι σωστός σε καμία από τις δυο περιπτώσεις*. Ενδεικτικά παραθέτουμε τις ακόλουθες δημοσιεύσεις, όπου και στις δυο περιπτώσεις ο τίτλος έχει αντικατασταθεί από το όνομα του συνεδρίου:

  ACM SIGMOD

  ACM SIGMOD 1999

  Στην περίπτωση αυτή ο χρήστης οφείλει να εξετάσει μόνο κατά πόσο αυτές οι δύο περιπτώσεις πράγματι αφορούν την ίδια δημοσίευση, ανεξάρτητα από το αν ο τίτλος τους είναι σωστός ή όχι. Εφόσον ο αλγόριθμος μας δεν εντοπίσει εν τέλει από μόνος του τη σωστή βιβλιογραφική αναφορά στην οποία αντιστοιχούν, ο χρήστης οφείλει και θα έχει την ευκαιρία να τις "συγχωνεύσει" με τις σωστές δημοσιεύσεις στο επόμενο παράθυρο, όπου εμφανίζονται συνολικά όλες οι δημοσιεύσεις που προέκυψαν από την αναζήτηση. Γενικότερα, υποχρέωση του χρήστη όταν επεξεργάζεται τα δεδομένα του παραθύρου επιβεβαίωσης είναι να αποφανθεί αποκλειστικά για το κατά πόσο οι εμφανιζόμενες δημοσιεύσεις πράγματι ταυτίζονται. Οποιαδήποτε άλλη επεξεργασία είτε γίνεται αυτόματα είτε έχει την ευκαιρία να την υλοποιήσει ο χρήστης στη συνέχεια.
- *Η "συγχώνευση" δυο δημοσιεύσεων είναι μη αναστρέψιμη διαδικασία*. Δηλαδή από τη στιγμή που επιβεβαιωθεί και συνδυαστούν οι πληροφορίες που περιέχουν οι δυο αυτές δημοσιεύσεις, δεν υπάρχει η δυνατότητα να διαχωριστούν. Αυτό οφείλεται στο γεγονός ότι κατά τη συγχώνευση δεν συγκρατείται πληροφορία με βάση την οποία να καθίσταται προφανές ποια από τις δυο δημοσιεύσεις συνείσφερε την πληροφορία σε ένα πεδίο της τελικής δημοσίευσης. Σκοπός μας είναι φυσικά να προστεθεί στο μέλλον και αυτή η δυνατότητα στο σύστημα.
- *Η "συγχώνευση" δυο δημοσιεύσεων δεν είναι αντιμεταθετική*. Πιο συγκεκριμένα, η συγχώνευση μιας δημοσίευσης β στην α δεν έχει το ίδιο αποτέλεσμα με τη συγχώνευση της α στη β. Αυτό οφείλεται κυρίως στο ότι κατά τη συγχώνευση συνδυάζονται όλα τα πεδία δυο δημοσιεύσεων με εξαίρεση τους συγγραφείς. Ως συγγραφείς της τελικής δημοσίευσης θεωρούνται οι συγγραφείς της "σωστής" δημοσίευσης, δηλαδή αυτής η

οποία "δέχεται" την άλλη δημοσίευση. Η παρατήρηση αυτή αφορά περισσότερο το επόμενο βήμα, κατά το οποίο ο χρήστης έχει τη δυνατότητα να μεταβάλλει τα αποτελέσματα του αλγορίθμου citation matching.

Αφού ολοκληρωθεί το βήμα της επιβεβαίωσης των συγχωνεύσεων, *το τελικό σύνολο των διαφορετικών δημοσιεύσεων που προκύπτει παρουσιάζεται στο χρήστη*, με τις νέες δημοσιεύσεις διαφορετικά χρωματισμένες από τις ήδη καταχωρημένες στη ΒΔ. Επίσης παρέχεται στο χρήστη η *δυνατότητα να μεταβάλλει τα αποτελέσματα*, συγχωνεύοντας δημοσιεύσεις που δεν εντοπίστηκαν από το σύστημα ή διαγράφοντας απροσδιόριστες και λανθασμένες δημοσιεύσεις.

Πιο συγκεκριμένα, επιλέγοντας τον τίτλο μιας δημοσίευσης, εμφανίζεται ένα νέο παράθυρο διαλόγου που δίνει τις εξής δυνατότητες στο χρήστη :

- να απορρίψει τη συγκεκριμένη δημοσίευση,
- να συγχωνεύσει μια λανθασμένη δημοσίευση με την αντίστοιχη σωστή, επιλέγοντας τον αύξοντα αριθμό της σωστής δημοσίευσης

Επίσης είναι διαθέσιμα και στην περίπτωση αυτή τα URL τόσο για το πλήρες κείμενο της δημοσίευσης όσο και για τα citations της που έχει εντοπίσει το Google Scholar. Έτσι, ο χρήστης έχει τη δυνατότητα να καταλήξει σε *ασφαλή συμπεράσματα*.

Όταν πλέον η επεξεργασία των δημοσιεύσεων έχει ολοκληρωθεί και το σύνολο των δημοσιεύσεων έχει λάβει την ορθή του μορφή, ο χρήστης πρέπει να πατήσει το κουμπί για την καταχώρηση των δημοσιεύσεων στη βάση δεδομένων. Για τις μεν νέες δημοσιεύσεις, προβλέπεται η εισαγωγή των πληροφοριών που αφορούν τους πίνακες paper, document_uri και title_variation, ενώ για τις ήδη καταχωρημένες δημοσιεύσεις προβλέπεται ενημέρωση των πληροφοριών των πινάκων αυτών. Έτσι, διατηρείται οποιαδήποτε νέα πληροφορία. Παρατηρούμε ότι δεν καταχωρούνται δεδομένα όσον αφορά τους συγγραφείς των δημοσιεύσεων, καθώς οι σχετικές πληροφορίες προκύπτουν μετά την ολοκλήρωση των επόμενων σταδίων.

3.5 Επίλυση των προβλημάτων mixed citation και split citation

Στόχος του σταδίου αυτού είναι *ο εντοπισμός των διαφορετικών επιστημόνων που συμμετέχουν στις δημοσιεύσεις που είναι καταχωρημένες στη βάση δεδομένων*, αυτών δηλαδή που ανακτήθηκαν από την τελευταία αναζήτηση από το Web αλλά και των ήδη καταχωρημένων πριν από αυτήν την αναζήτηση. Για την επίτευξη αυτού του στόχου

χρειάζεται να αντιμετωπιστούν τα προβλήματα mixed citation, κατά το οποίο δυο διαφορετικοί επιστήμονες έχουν τα ίδια ονόματα, και split citation, κατά το οποίο ο ίδιος επιστήμονας εμφανίζεται με διαφορετικά ονόματα. Πρέπει να επισημανθεί ότι απαραίτητη προϋπόθεση για την επιτυχημένη επίλυση των προβλημάτων αυτών είναι η σωστή εκτέλεση των δυο προηγούμενων σταδίων, δηλαδή του citation matching. Σε αντίθετη περίπτωση οι παραλλαγές μιας δημοσίευσης που δεν εντοπίστηκαν οδηγούν στην αλλοίωση των αποτελεσμάτων της συγκεκριμένης διαδικασίας καθώς οι συγγραφείς της συγκεκριμένης δημοσίευσης εμφανίζονται περισσότερες από μια φορές.

Η προσέγγιση μας στα προβλήματα mixed citation και split citation μοιάζει με τις αντίστοιχες στη βιβλιογραφία, μόνο όσον αφορά στα δεδομένα που χρησιμοποιούμε για τη σύγκριση δυο επιστημόνων. Για την ακρίβεια, κατά τη σύγκριση αυτή εξετάζουμε την ομοιότητα των εξής πεδίων:

- ονόματα επιστημόνων,
- ονόματα των συν-συγγραφέων τους,
- τίτλοι των δημοσιεύσεων,
- URL των δημοσιεύσεων,

Ωστόσο, δε μπορέσαμε να χρησιμοποιήσουμε κανένα από τους δοκιμασμένους στη βιβλιογραφία αλγορίθμους, λόγω της ιδιαιτερότητας με την οποία εμφανίζονται τα προβλήματα αυτά στην περίπτωσή μας. Πιο συγκεκριμένα, οι αλγόριθμοι της βιβλιογραφίας εφαρμόζονταν στην πλειονότητά τους σε ένα περιορισμένο, προσαρμοσμένο στις ανάγκες των πειραμάτων, σύνολο δεδομένων. Σε όλες, δηλαδή, τις σχετικές προσπάθειες γινόταν αρχικά επιλογή ενός συνόλου δημοσιεύσεων στις οποίες συμμετείχε ένας συγγραφέας με συνηθισμένο επίθετο (π.χ. Johnson) και στόχος του αλγόριθμου ήταν να διαχωρίσει τους συγγραφείς με το συγκεκριμένο επίθετο. Δεν επιχειρούσαν δηλαδή να βρουν όλους τους διαφορετικούς συγγραφείς που συμμετείχαν στις επιλεγμένες δημοσιεύσεις, αλλά εξέταζαν την επίδοση του αλγορίθμου σχετικά με το συγκεκριμένο επίθετο. Αντίθετα, το πεδίο εφαρμογής του αλγορίθμου μας είναι μια πραγματικού χρόνου εφαρμογή που πρέπει να ανταποκρίνεται επιτυχώς σε οποιαδήποτε δεδομένα εισόδου. Συνεπώς, χρειαζόμαστε στην περίπτωσή μας έναν αλγόριθμο ο οποίος να λαμβάνει υπόψη του τα εξής :

- *Οι συγγραφείς που συνεργάζονται σε μια δημοσίευση είναι σίγουρα διαφορετικά πρόσωπα.* Αν δεν εκμεταλλευτούμε το γεγονός αυτό και επιχειρήσουμε να συγκρίνουμε συγγραφείς που συμμετέχουν στην ίδια δημοσίευση, το πιθανότερο είναι ότι η μετρική μας θα βρει ότι ταυτίζονται. Άλλωστε οι επιστήμονες αυτοί έχουν κοινές όλες τις ιδιότητες που αυτή συγκρίνει (τίτλος και URL δημοσίευσης), με εξαίρεση τη λίστα των συν-συγγραφέων, η οποία όμως διαφέρει σε ένα μόνο όνομα.

- *Ένας συγγραφέας μπορεί να ταυτίζεται το πολύ με έναν από τους συγγραφείς μιας δημοσίευσης*. Αν αγνοήσουμε το γεγονός αυτό, προκύπτει η ακόλουθη διαστρέβλωση των αποτελεσμάτων:

  Αν βρεθεί ότι ο δοσμένος συγγραφέας ταυτίζεται με έναν από τους συγγραφείς μιας δημοσίευσης, το πιθανότερο είναι να υποδείξει η μετρική μας ότι και οι υπόλοιποι συγγραφείς της ταυτίζονται με τον δοσμένο. Και σε αυτή την περίπτωση το πρόβλημα δημιουργείται από το γεγονός ότι οι συνεργαζόμενοι σε μια δημοσίευση συγγραφείς έχουν ταυτόσημα σύνολα χαρακτηριστικών (τίτλος και URL δημοσίευσης), με εξαίρεση την λίστα των συν-συγγραφέων που διαφέρει σε όνομα.

  Για το λόγο αυτό, όταν αναζητούμε με ποιον από τους συγγραφείς μιας δημοσίευσης είναι πιθανότερο να ταυτίζεται ένας δοσμένος συγγραφέας, βρίσκουμε πρώτα αυτόν που τα ονόματά τους είναι περισσότερο όμοια (με μια μετρική της κατηγορίας edit distance-αναπαράσταση με χαρακτήρες) και εκτελούμε τις υπόλοιπες συγκρίσεις μόνο με αυτόν.
- Είναι άσκοπο και ασύμφορο από άποψη χρόνου και μνήμης να επαναλαμβάνεται ο αλγόριθμος για δεδομένα που είναι ήδη καταχωρημένα στη βάση δεδομένων. Πρέπει δηλαδή ο αλγόριθμός μας *να ξεκινάει τη λειτουργία βασιζόμενος στα αποτελέσματα στα οποία είχε καταλήξει τις προηγούμενες φορές που εφαρμόστηκε*.

Επιπλέον, ο αλγόριθμος ομαδοποίησης που θα εφαρμόσουμε πρέπει απαραιτήτως να έχει τα ακόλουθα χαρακτηριστικά :

- *max intra-cluster similarity*: να επιτυγχάνει τη μέγιστη δυνατή ομοιότητα μεταξύ των συγγραφέων που ανήκουν στην ίδια ομάδα-cluster.
- *partitioning*: να διαχωρίζει το αρχικό σύνολο συγγραφέων σε *ξένα* μεταξύ τους υποσύνολα. Πρέπει δηλαδή να μην επιτρέπει επικαλύψεις μεταξύ των ομάδων-clusters.
- *self-determined number of clusters*: το πλήθος των ομάδων πρέπει να καθορίζεται με βάση τα εκάστοτε δεδομένα, δηλαδή να μην είναι προκαθορισμένο.

Στο σημείο αυτό πρέπει να διευκρινίσουμε ότι *το αρχικό σύνολο των συγγραφέων αποτελείται από <u>όλους</u> τους συγγραφείς <u>όλων</u> των καταχωρημένων στη βάση δημοσιεύσεων*. Δηλαδή, το σύνολο αυτό αποτελείται από ονόματα, κάθε ένα από τα οποία είναι συνδεδεμένο με ένα διάνυσμα χαρακτηριστικών που το απαρτίζουν οι συν-συγγραφείς, ο τίτλος και τα URL της δημοσίευσης από την οποία προήλθε. Οι συγγραφείς αυτοί αποτελούν τους κόμβους ενός μη κατευθυνόμενου γράφου. Στόχος του αλγορίθμου μας είναι να συνδέσει με ακμές όσους κόμβους-συγγραφείς πιστεύει ότι ταυτίζονται και στη συνέχεια να διαχωρίσει το γράφο σε συνεκτικές συνιστώσες (connected components). Μετά την ολοκλήρωση της διαδικασίας, κάθε συνεκτική συνιστώσα θα αντιστοιχεί σε ένα πραγματικό επιστήμονα και περιέχει τις πληροφορίες που χρειάζεται για να συμπληρωθούν οι πίνακες scholar, related_names και contributes_to. <u>Κάθε ομάδα-cluster πρέπει δηλαδή να περιλαμβάνει</u> :

- *το πιο αντιπροσωπευτικό όνομα* του αντίστοιχου συγγραφέα (πίνακας scholar). Ως τέτοιο θεωρούμε το μεγαλύτερο σε μήκος συνώνυμό του, με το σκεπτικό ότι περιέχει περισσότερη πληροφορία, πλήρες όνομα για παράδειγμα.
- τα *συνώνυμα* που αντιστοιχούν σε κάθε συγγραφέα (πίνακας related_names). Τα συνώνυμα αυτά είναι στην ουσία *όλα τα διαφορετικά* ονόματα που περιέχει κάθε cluster.
- τις *δημοσιεύσεις* στις οποίες συμμετέχει ο συγγραφέας αλλά και τη *σειρά* με την οποία εμφανίζεται το όνομά του σε αυτές (πίνακας contributes_to). Επιπλέον, πρέπει να συγκρατείται και το *συνώνυμο* με το οποίο εμφανίζεται στις δημοσιεύσεις αυτές.

Λαμβάνοντας αυτά υπόψη, <u>τα βήματα του αλγόριθμου</u> που αναπτύξαμε διαμορφώνονται ως εξής :

- Αρχικά, ανακαλούνται από τη βάση δεδομένων *όλες* οι καταχωρημένες δημοσιεύσεις, γνωρίζοντας όμως ποιες από αυτές ήταν ήδη καταχωρημένες και ποιες προέκυψαν από την τελευταία αναζήτηση. Τονίζουμε το όλες γιατί κάτι τέτοιο είναι αναπόφευκτο, εφόσον είναι πολύ πιθανό κάποιοι από τους συγγραφείς των νέων δημοσιεύσεων να συμμετείχαν σε εργασία που ήταν ήδη καταχωρημένη στη ΒΔ. Είναι προφανώς απαραίτητο να γίνονται τέτοιες ταυτίσεις κάτι που δεν μπορεί να γίνει αν δεν εξεταστούν όλες οι καταχωρημένες στη ΒΔ δημοσιεύσεις.
- Δημιουργούμε ένα *μη-κατευθυνόμενο γράφο* με τη βοήθεια της βιβλιοθήκης JUNG. Στον γράφο αυτό τοποθετούμε έναν κόμβο για κάθε έναν από τους συγγραφείς κάθε δημοσίευσης, δημιουργούμε δηλαδή το αρχικό σύνολο συγγραφέων. Οι κόμβοι αυτοί είναι αντικείμενα της κλάσης AuthorVertex (βλέπε υποενότητα 4.3.4), έτσι ώστε να διατηρούν όλη την πληροφορία που σχετίζεται με τη συμμετοχή κάθε συγγραφέα στη δημοσίευση από την οποία προήλθε.

  Για τους συγγραφείς κάθε δημοσίευσης που ήταν <u>ήδη καταχωρημένη</u> στη βάση, δηλαδή για τους οποίους είχαν ήδη λυθεί τα προβλήματα mixed και split citation, τοποθετούμε ακμές μεταξύ αυτών που έχουν το ίδιο ScholarId. Με τον τρόπο αυτό αποφεύγουμε την επανάληψη της επεξεργασίας για τα δεδομένα που ήταν αποθηκευμένα στη ΒΔ πριν την τελευταία αναζήτηση από το Web.

  Με τον τρόπο αυτό, τοποθετούμε *<u>σε μια σειρά</u> τους κόμβους του γράφου που αντιστοιχούν στον ίδιο επιστήμονα, όπως τουλάχιστον αυτός είχε προκύψει από προηγούμενη επεξεργασία*. Οι κόμβοι αυτοί συνδέονται, επομένως, μεταξύ τους με ένα μονοπάτι. Σαν αποτέλεσμα, όταν ο γράφος διασπαστεί σε connected components, θα τοποθετηθούν στην ίδια ομάδα. Έτσι διατηρείται όση πληροφορία έχει προκύψει από επεξεργασία προηγούμενων on-line αναζητήσεων.
- Το επόμενο βήμα είναι *η τοποθέτηση ακμών που αφορούν τους συγγραφείς των νέων δημοσιεύσεων*, ακμών δηλαδή που στο ένα άκρο τους θα έχουν επιστήμονα νέας εργασίας. Για την τοποθέτηση των ακμών γίνεται η εξής επεξεργασία :

Όπως και στην επίλυση του προβλήματος citation matching (ενότητα 3.3), *οι δημοσιεύσεις ταξινομούνται* έτσι ώστε να προηγούνται οι καταχωρημένες και να ακολουθούν οι νέες με φθίνουσα σειρά ως προς των αριθμό των citations και των URL.

Για κάθε νέα δημοσίευση και για κάθε συγγραφέα της, εξετάζουμε όλες τις προηγούμενές της για να βρούμε τον πιο όμοιο του συγγραφέα. Στις συγκρίσεις αυτές, οι οποίες στην ουσία αποτελούν ένα είδος φιλτραρίσματος ώστε να αποφευχθούν περιττές συγκρίσεις με όλους τους συγγραφείς, χρησιμοποιούμε πάλι τη *μετρική Jaro*.

Αν η ομοιότητα του τρέχοντα συγγραφέα με τον πιο όμοιό του δεν υπερβαίνει το 0.8, θεωρούμε ότι δεν έχει βρεθεί κανένας πιθανά ταυτόσημος συγγραφέας.

Διαφορετικά, συγκρίνουμε και τα διανύσματα των χαρακτηριστικών που συνδέονται με τους συγγραφείς αυτούς για να διαπιστώσουμε αν πράγματι οι συγγραφείς αυτοί ταυτίζονται. Η *σύγκριση των διανυσμάτων χαρακτηριστικών* που συνδέονται με κάθε επιστήμονα υλοποιείται ως εξής:

1. Για κάθε έναν από τους συγκρινόμενους επιστήμονες, σχηματίζουμε μια συμβολοσειρά με τα ονόματα των συγγραφέων που συνεργάστηκαν μαζί του στις τρέχουσες δημοσιεύσεις. Τις συμβολοσειρές αυτές τις συγκρίνουμε με μια μετρική SoftTFIDF σε συνδυασμό με τη Jaro και κατώφλι 0.8. Η τιμή που προκύπτει από τη σύγκριση αυτή υποδηλώνει το *ποσοστό των κοινών συν-συγγραφέων*.

2. Συγκρίνουμε τους τίτλους των τρεχουσών δημοσιεύσεων, πάλι με την ίδια μετρική, ώστε να πάρουμε ένα *ποσοστό των κοινών λέξεων*. Σε περίπτωση που είτε μια είτε και οι δυο δημοσιεύσεις δεν έχουν ένα μοναδικό τίτλο, αλλά διάφορες παραλλαγές του λόγω του προβλήματος citation matching, η τιμή που υπολογίζουμε είναι ο μέσος όρος που προκύπτει από τη διαδοχική σύγκριση των διαφορετικών τίτλων.

3. Σχηματίζουμε για κάθε έναν από τους συγγραφείς μια συμβολοσειρά που αποτελείται *από τα domains όλων των URLs της αντίστοιχης δημοσίευσης*. Σημαντικό είναι ότι διατηρούμε μόνο τα domains, για παράδειγμα από το *http://citeseer.ist.psu.edu/522320.html* κρατάμε μόνο το *citeseer.ist.psu.ed,* γιατί διαφορετικά το αποτέλεσμα αλλοιώνεται. Για την ακρίβεια, το πρωτόκολλο στην αρχή του URL (*http:///*) αυξάνει την ομοιότητα καθώς στη συντριπτική πλειονότητα των URL είναι το http. Αντίθετα,, το σχετικό path (*/522320.html*) κατά κανόνα μειώνει την ομοιότητα, αφού δεν είναι δυνατόν δυο διαφορετικές δημοσιεύσεις να έχουν το ίδιο σχετικό path, δηλαδή να αντιστοιχούν στο ίδιο έγγραφο.

Η σύγκριση των domains των URLs στόχο έχει να μας δώσει *ένα μέτρο της ομοιότητας είτε των επιστημονικών περιοχών με τις οποίες ασχολούνται οι συγκρινόμενοι συγγραφείς* (π.χ. το domain www.acm.com υποδηλώνει την επιστήμη των υπολογιστών) *είτε των ιδρυμάτων στα οποία εργάζονται αυτοί* (π.χ. αν και οι δυο συγκρινόμενες δημοσιεύσεις

περιλαμβάνουν το domain www.ntua.gr κατά πάσα πιθανότητα οι συγγραφείς τους εργάζονται στο ΕΜΠ).

4. Αθροίζουμε την ομοιότητα των ονομάτων των δυο επιστημόνων με τις τιμές που προκύπτουν από τις προηγούμενες συγκρίσεις θέτοντας σε όλες τις τιμές εκτός από το ποσοστό των κοινών συν-συγγραφέων βάρος 1. Στην ομοιότητα των συν-συγγραφέων αποδίδουμε βάρος 1.4, καθώς όπως τονίζεται και στη βιβλιογραφία, αυτή είναι η καθοριστικότερη παράμετρος για την ταύτιση δυο επιστημόνων. Το συνολικό άθροισμα το συγκρίνουμε με ένα κατώφλι που έχουμε θέσει στο 1.9. Αν δηλαδή το παραπάνω άθροισμα ξεπεράσει την τιμή αυτή, θεωρούμε ότι οι συγγραφείς ταυτίζονται και τοποθετούμε στο γράφο μια ακμή που συνδέει τους αντίστοιχους κόμβους.

Στο σημείο αυτό πρέπει να παρατηρήσουμε τα εξής για το συγκεκριμένο βήμα του αλγορίθμου :

- *Ιδιαίτερα κρίσιμη είναι η τιμή του κατωφλιού του φιλτραρίσματος*. Με το όρο αυτό αναφερόμαστε στην ελάχιστη τιμή ομοιότητας που επιστρέφει η μετρική Jaro μεταξύ του τρέχοντος ονόματος και του πιο όμοιού του συγγραφέα από την υπό εξέταση δημοσίευση.

  Αν η τιμή του κατωφλιού αυτού είναι μεγάλη, είναι εξαιρετικά πιθανό ονόματα συγγραφέων που αντιστοιχούν στον ίδιο επιστήμονα να θεωρούνται διαφορετικά. Δε θα συγκρίνεται δηλαδή το διάνυσμα των χαρακτηριστικών τους επειδή τα ονόματά τους διαφέρουν αρκετά.

  Αντίθετα, αν η τιμή του κατωφλιού είναι μικρή, είναι μεγαλύτερη η πιθανότητα να ταυτιστούν δυο προφανώς διαφορετικοί συγγραφείς, επειδή έτυχε να έχουν αρκετά όμοιο διάνυσμα χαρακτηριστικών.

  Μετά από πειραματισμούς, επιλέξαμε να θέσουμε το κατώφλι φιλτραρίσματος στην τιμή 0.8, μια σχετικά ψηλή τιμή για τη μετρική Jaro που όμως επιτρέπει σημαντικές παραλλαγές στα ονόματα. Στόχος μας είναι να μειώσουμε τις περιπτώσεις που ταυτίζονται διαφορετικοί συγγραφείς, αυξάνοντας βέβαια τις περιπτώσεις που συνώνυμα του ίδιου συγγραφέα θεωρούνται διαφορετικοί επιστήμονες. Αυτό γίνεται επειδή, όπως θα φανεί παρακάτω, είναι ευκολότερο για το χρήστη να εντοπίσει την δεύτερη περίπτωση κατά την επιβεβαίωση των αποτελεσμάτων. Αντίθετα, απαιτείται ιδιαίτερη προσοχή και σχολαστικότητα για τον εντοπισμό της πρώτης περίπτωσης.
- *Ενδέχεται μια δημοσίευση να έχει γραφτεί από ένα μόνο συγγραφέα ή να μην έχει βρεθεί κάποιο URL για αυτήν*. Στις περιπτώσεις αυτές η σύγκριση των συν-συγγραφέων και των URL δεν θα δώσει αποτελέσματα. Είναι απαραίτητο επομένως σε τέτοιες περιπτώσεις να μειώνουμε το κατώφλι, για να μην αγνοήσουμε περιπτώσεις ταύτισης. Αναλυτικότερα, σε περίπτωση απουσίας συν-συγγραφέων, το

κατώφλι μειώνεται κατά 0.5. Αν τουλάχιστον μια δημοσίευση δεν έχει κανένα URL, το κατώφλι μειώνεται επίσης κατά 0.5. Σε περίπτωση που υπάρχει έλλειψη και για τα δυο χαρακτηριστικά, το κατώφλι μειώνεται συνολικά κατά 0.8 και διαμορφώνεται στο 1.1.

- Σε αντίθεση με τους συγγραφείς των καταχωρημένων δημοσιεύσεων, οι συγγραφείς των νέων δημοσιεύσεων συνδέονται με *όλους* τους επιστήμονες με τους οποίους ο αλγόριθμος υποδηλώνει ταίριασμα. Πιο συγκεκριμένα, κατά την ανάκληση των επιστημόνων που ήταν ήδη καταχωρημένοι στη βάση συνδέουμε τους αντίστοιχους κόμβους σε μια ευθεία γραμμή, δηλαδή με τον ελάχιστο δυνατό αριθμό ακμών ώστε να ανήκουν στο ίδιο connect component. Αντίθετα, οι συγγραφείς των νέων δημοσιεύσεων συνδέονται με όλους τους κόμβους του ίδιου επιστήμονα, δηλαδή με το μέγιστο αριθμό ακμών. Αυτό γίνεται για λόγους ευρωστίας του αλγορίθμου.

- Το τελευταίο βήμα του αλγορίθμου αυτού είναι *ο διαχωρισμός του γράφου σε connected components*. Η λειτουργία αυτή πραγματοποιείται με μια από τις συναρτήσεις χειρισμού των γράφων που διαθέτει η βιβλιοθήκη JUNG. Οι κόμβοι που αποτελούν ένα τέτοιο σύνολο αντιστοιχούν όλοι στον ίδιο επιστήμονα. Από την επεξεργασία κάθε τέτοιου συνόλου προκύπτει ένα αντικείμενο της κλάσης Scholar που μοντελοποιεί ένα επιστήμονα. Για την ακρίβεια, περιέχει το πιο αντιπροσωπευτικό του όνομα, τα συνώνυμά του αλλά και τις δημοσιεύσεις στις οποίες συμμετέχει καθώς επίσης και τη σειρά με την οποία εμφανίζεται το όνομά του. Τα αντικείμενα της κλάσης Scholar ταξινομούνται στη συνέχεια με αλφαβητική σειρά επιθέτου, ώστε να παρουσιαστούν κατάλληλα στο χρήστη, στο επόμενο στάδιο της διαδικασίας αυτής.

3.6 Επεξεργασία των διαφορετικών επιστημόνων από το χρήστη και καταχώρησή τους στη βάση δεδομένων

Στο τελευταίο αυτό στάδιο της διαδικασίας της on-line αναζήτησης, δίνεται η ευκαιρία στο χρήστη να *επεξεργαστεί και να μεταβάλλει τα αποτελέσματα του προηγούμενου σταδίου, που αφορούν τον εντοπισμό των διαφορετικών επιστημόνων*, πριν αυτά καταχωρηθούν στη βάση δεδομένων.

Έτσι, το πρώτο βήμα στο στάδιο αυτό είναι *η εμφάνιση των επιστημόνων στο χρήστη*. Κάθε συγγραφέας εμφανίζεται με το πιο αντιπροσωπευτικό του συνώνυμο, το πλήθος των συνωνύμων του και τον αριθμό των δημοσιεύσεων του, ενώ του έχει αποδοθεί και ένας αύξων αριθμός. Επίσης, ο χρήστης μπορεί να επιλέξει να εμφανίσει τα συνώνυμα σε ένα νέο

παράθυρο. Αντίστοιχα, μπορεί να επιλέξει να εμφανίσει νέο tab με τις δημοσιεύσεις του συγκεκριμένου συγγραφέα ταξινομημένες και ομαδοποιημένες με βάση τα συνώνυμά του. Μέσω αυτού του tab παρέχεται η δυνατότητα στο χρήστη να τροποποιήσει κατά το δοκούν τα αποτελέσματα του προηγούμενου αλγορίθμου.

Πιο συγκεκριμένα, οι αλλαγές που μπορεί να κάνει ο χρήστης είναι οι εξής :

- *Να διαγράψει ή να μεταφέρει <u>ένα συνώνυμο</u> σε ένα άλλο επιστήμονα.*

  Με τη δυνατότητα αυτή αντιμετωπίζονται οι περιπτώσεις που *το σύνολο των δημοσιεύσεων* που αντιστοιχούν σε ένα συνώνυμο έχει αποδοθεί λανθασμένα σε ένα επιστήμονα. Απαιτείται δηλαδή είτε η μεταφορά τους σε άλλον επιστήμονα είτε η διαγραφή τους. Εξίσου πιθανό είναι το συγκεκριμένο συνώνυμο να αντιστοιχεί σε ένα καινούριο επιστήμονα, που δεν προέκυψε δηλαδή από το clustering. Ο χρήστης έχει τη δυνατότητα να πραγματοποιήσει την απαιτούμενη αλλαγή επιλέγοντας το συνώνυμο που επιθυμεί να μεταβάλει. Στο παράθυρο διαλόγου που εμφανίζεται καθορίζει την επιθυμητή ενέργεια. Αξίζει να σημειωθεί ότι όλες αυτές οι περιπτώσεις μεταβολής των συνωνύμων ενός επιστήμονα αλλάζουν και το αντιπροσωπευτικό του συνώνυμο.

  Όσον αφορά τη *μεταφορά συνώνυμου σε άλλο επιστήμονα*, αυτή γίνεται έτσι ώστε αν ο νέος επιστήμονας δεν διαθέτει ήδη το συγκεκριμένο συνώνυμο, αυτό να προστίθεται στο σύνολο των συνωνύμων του. Αντίθετα, αν συμπεριλαμβάνεται ήδη στα συνώνυμά του, οι αντίστοιχες εργασίες απλά προστίθενται σε αυτές του ήδη υπάρχοντος συνωνύμου.

  Η επιλογή της *διαγραφής συνωνύμου* φαίνεται εκ πρώτης όψεως υπερβολική και περιττή. Λαμβάνοντας όμως υπόψη την χαμηλή ποιότητα των δεδομένων που έχουν αρκετές βιβλιογραφικές αναφορές του Google Scholar, κρίνεται απαραίτητη. Ενδεικτικά παραθέτουμε μια εγγραφή BibTex όπου κάποιοι από τους συγγραφείς της δημοσίευσης εμφανίζονται δυο φορές. Ο αλγόριθμός μας αναπόφευκτα τους θεωρεί διαφορετικούς επιστήμονες, καθώς δε συγκρίνει μεταξύ τους συγγραφείς που συνεργάζονται σε μια δημοσίευση. Συχνό είναι επίσης το φαινόμενο λέξεις του τίτλου να εμφανίζονται σαν ονόματα συγγραφέων και στη συνέχεια να παρουσιάζονται σαν επιστήμονες.

```
@article{adamic46hhl,
title={{href $\{$http://www. liafa. jussieu. fr/\~{} fabien/generation$\}$}},
author={Adamic, LA and Lukose, RM and Puniyani, AR and Adamic, LA and
Lukose, RM and Huberman, BA and Aiello, W. and Chung, F. and Lu, L. and
Berge, C. and others},
journal={Proc. of the 30th $\{$ACM$\}$$\{$STOC$\}$},
volume={46},
pages={79--89},
publisher={ACM Press}
}
```

- *Να διαγράψει ή να μεταφέρει <u>μια δημοσίευση</u> σε ένα άλλο επιστήμονα.*

Μέσω αυτής της δυνατότητας επιχειρείται να δοθεί λύση στις περιπτώσεις όπου *μεμονωμένες εργασίες* έχουν αποδοθεί λανθασμένα σε ένα επιστήμονα. Σε τέτοιες περιπτώσεις, ο χρήστης πρέπει να είναι σε θέση είτε να δημιουργήσει ένα νέο επιστήμονα με βάση τις συγκεκριμένες δημοσιεύσεις είτε να τις μεταφέρει σε άλλο επιστήμονα ή ακόμα και να τις διαγράψει. Αυτό επιτυγχάνεται κάνοντας κλικ πάνω στον τίτλο της λανθασμένης εργασίας για να εμφανιστεί το σχετικό παράθυρο διαλόγου, από όπου ο χρήστης επιλέγει την επιθυμητή ενέργεια.

Και στην περίπτωση αυτή η *μεταφορά δημοσίευσης* αυτή γίνεται αυτόματα. Έτσι, αν ο νέος επιστήμονας δεν διαθέτει το αντίστοιχο συνώνυμο, αυτό προστίθεται στο σύνολο των συνωνύμων του. Αντίθετα, αν το διαθέτει, η δημοσίευση απλά προστίθεται σε αυτό. Επιπλέον, στην περίπτωση διαγραφής μιας εργασίας εξετάζεται το ενδεχόμενο η δημοσίευση αυτή ήταν η μοναδική του συγκεκριμένου συνωνύμου του επιστήμονα, οπότε αυτό πρέπει να διαγραφεί.

Πρέπει να σημειώσουμε στο σημείο αυτό ότι σε περίπτωση που μετά τις αλλαγές του χρήστη ένας επιστήμονας μείνει χωρίς καμία δημοσίευση, δεν διαγράφεται αμέσως από το σύνολο των συγγραφέων. Απλά το σύστημα φροντίζει να μην καταχωρηθεί στη βάση δεδομένων, όταν ο χρήστης ολοκληρώσει την επεξεργασία.

3.7 Αναζήτηση citations για μια δημοσίευση

Μια αναζήτηση από το web ολοκληρώνεται όταν ο χρήστης τελειώσει με την επεξεργασία των επιστημόνων και επιλέξει την καταχώρησή τους στη βάση δεδομένων. Μόνο τότε επιτρέπεται να ξεκινήσει νέα on-line αναζήτηση, είτε αυτή αφορά τις εργασίες ενός άλλου επιστήμονα είτε τα citations κάποιας από τις νέες δημοσιεύσεις. Επιλέξαμε δηλαδή η εφαρμογή να μην επιτρέπει την ταυτόχρονη εκτέλεση δυο αναζητήσεων, γιατί διαφορετικά θα ήταν ιδιαίτερα περίπλοκη η επεξεργασία της ορθότητας των αποτελεσμάτων τους.

Ειδικότερα, η αναζήτηση των citations θα μπορούσαμε να γίνεται αυτόματα κατά τη διάρκεια της αναζήτησης των δημοσιεύσεων ενός συγκεκριμένου συγγραφέα και μάλιστα για όλες τις δημοσιεύσεις που του αντιστοιχούν. Ωστόσο κάτι τέτοιο θα είχε ως αποτέλεσμα να απαιτείται από το χρήστη να επικυρώσει ένα τεράστιο όγκο δεδομένων, τόσο όσον αφορά το citation matching όσο για την εύρεση των διαφορετικών επιστημόνων. Άλλωστε, συνήθως το πλήθος των citations ενός επιστήμονα υπερβαίνει τον αριθμό των δημοσιεύσεών του. Ομοίως, και ο αριθμός των citing authors είναι πολλαπλάσιος του πλήθους των συν-συγγραφέων του.

Για τους παραπάνω λόγους, επιλέξαμε να εκτελείται η on-line αναζήτηση των citations ξεχωριστά για κάθε εργασία ενός επιστήμονα ως εξής : όταν ο χρήστης πατήσει το αντίστοιχο

κουμπί, εμφανίζεται ένα νέο tab στο κεντρικό παράθυρο της εφαρμογής και ξεκινάει αυτόματα η επιθυμητή αναζήτηση. Η εξέλιξή της είναι πανομοιότυπη με την αναζήτηση των εργασιών ενός επιστήμονα. Αφότου δηλαδή ανακτηθούν όλες οι βιβλιογραφικές αναφορές, ο χρήστης καλείται να επιβεβαιώσει τα αποτελέσματα του citation matching αλγόριθμου. Στην συνέχεια, πρέπει να "συγχωνεύσει" χειρωνακτικά αυτές που ταυτίζονται αλλά δεν εντοπίστηκαν από τον προηγούμενο αλγόριθμο. Σειρά έχει η αναγνώριση των διαφορετικών επιστημόνων από τον αντίστοιχο αλγόριθμο, η επικύρωση των αποτελεσμάτων του από το χρήστη και τέλος η καταχώρησή τους στη βάση δεδομένων. Ακολούθως, μπορεί να γίνει μια νέα αναζήτηση στο Web.

3.8 Αξιολόγηση Αλγορίθμων

Στην ενότητα αυτή θα παρουσιάσουμε κάποιες αρχικές μετρήσεις που έγιναν για να εκτιμηθεί η επίδοση των αλγορίθμων που αναπτύξαμε για την επίλυση των προβλημάτων citation matching, mixed και split citation. Αξίζει να τονιστεί ότι τα αποτελέσματα των μετρήσεων αυτών είναι ενδεικτικά και όχι αντιπροσωπευτικά της πραγματικής επίδοσης, καθώς δεν πρόκειται για μια διεξοδική αξιολόγηση. Κάτι τέτοιο θα απαιτούσε ένα προσεκτικά επιλεγμένο σύνολο δεδομένων στο οποίο θα αντιπροσωπεύονταν *διάφορες επιστημονικές περιοχές και εθνικότητες*. Οι συγκεκριμένοι παράγοντες είναι καθοριστικοί για την επίδοση ενός αλγορίθμου citation matching αλλά κυρίως ενός αλγορίθμου name disambiguation (δηλαδή mixed και split citation) για τους παρακάτω λόγους.

- Ο επιστημονικός κλάδος αποτελεί ένα παράγοντα που επηρεάζει την ποιότητα των εγγραφών του Google Scholar, καθώς κάποιοι κλάδοι, όπως η επιστήμη των υπολογιστών, έχουν εγγραφές με κατά μέσο όρο πιο σωστά δεδομένα από τις αντίστοιχες για παράδειγμα της κοινωνιολογίας. Σημαντικές μεταβολές παρατηρούνται επίσης και στο μέσο αριθμό συγγραφέων μιας δημοσίευσης στους διάφορους επιστημονικούς κλάδους. Για παράδειγμα έχει αποδειχτεί ότι στις δημοσιεύσεις που αφορούν τα οικονομικά συμμετέχουν εν γένει λιγότεροι από ότι σε μια εργασία για την ιατρική. Λαμβάνοντας υπόψη ότι οι συν-συγγραφείς είναι το καθοριστικότερο κριτήριο για την ταύτιση επιστημόνων κατανοούμε πόσο σημαντική είναι η επιρροή του επιστημονικού κλάδου στην επίδοση των αλγορίθμων μας.
- Εξίσου σημαντική, αν όχι σημαντικότερη, είναι και η επίδραση της εθνικότητας των συγγραφέων στην επίδοση ενός αλγορίθμου για τα προβλήματα mixed και split citation λόγω της διαφορετικής ποικιλίας των επιθέτων κάθε εθνικότητας. Πιο συγκεκριμένα, κάθε επιστήμονας αντιπροσωπεύεται από το αρχικό γράμμα του ονόματός και

ολόκληρο το επίθετό του. Αυτό έχει ως αποτέλεσμα επιστήμονες με καταγωγή από χώρες όπως η Κίνα ή η Κορέα, δηλαδή χώρες με μεγάλο πληθυσμό και περιορισμένο αριθμό διαφορετικών επιθέτων, να είναι εξαιρετικά δύσκολο να διαχωριστούν. Σε αυτό πρέπει να συνυπολογιστεί και το γεγονός ότι τα επίθετά τους αποτελούνται ως επί το πλείστον από 2 έως 4 γράμματα, δυσχεραίνοντας έτσι το έργο των μετρικών σύγκρισης ονομάτων, όπως η Jaro.

Λαμβάνοντας αυτά υπόψη επιλέξαμε να επιχειρήσουμε μια ενδεικτική αξιολόγηση με ένα περιορισμένο σύνολο δεδομένων. Για την ακρίβεια, επιλέξαμε το σύνολο αυτό να αποτελείται από τις δημοσιεύσεις και τα citations που επιστρέφει το Google Scholar για το όνομα επιστήμονα G. Paliouras και *πράγματι* αντιστοιχούν στο ερευνητή του Ε.Κ.Ε.Φ.Ε. "Δημόκριτος" Γεώργιο Παλίουρα. Τα αποτελέσματα των μετρήσεων αυτών παρουσιάζονται στο ακόλουθο πίνακα:

Αλγόριθμος	Προτεινόμενα Ταιριάσματα	Λάθος Ταιριάσματα	Πραγματικά ταιριάσματα	Μη Ταιριάσματα
Citation Matching	70	22	56	8
Name Disambiguation	732	40	719	27

Εκφρασμένα σε μετρικές της εξαγωγής πληροφορίας τα παραπάνω αποτελέσματα διαμορφώνονται ως εξής:

Αλγόριθμος	Ακρίβεια	Ανάκληση
Citation Matching	68.57%	85.71%
Name Disambiguation	94.54%	96.24%

Στο σημείο αυτό πρέπει να επισημάνουμε ότι η **ακρίβεια** (*precision*) εκφράζει το ποσοστό των ζευγών που ορθά αναγνωρίστηκαν ως ταιριάσματα, ενώ η **ανάκληση** (*recall*) αποτελεί το ποσοστό των ζευγών που ήταν πράγματι ταιριάσματα και αναγνωρίστηκαν ως τέτοια.

Τα αποτελέσματα αυτά, αν και μάλλον ικανοποιητικά, σε καμία περίπτωση δεν πρέπει να θεωρηθούν αρκετά γενικά και αντιπροσωπευτικά ώστε να μας οδηγήσουν σε ασφαλή συμπεράσματα. Αντίθετα, οι παρακάτω λόγοι μας αναγκάζουν να τα δεχθούμε απλά ως μια πρώτη ένδειξη της επίδοσης των αλγορίθμων.

- Το δείγμα για τον αλγόριθμο citation matching είναι εξαιρετικά περιορισμένο.
- Τα αποτελέσματα έχουν αλλοιωθεί από την χαμηλή ποιότητα πολλών από τις εγγραφές που επιστρέφει το Google Scholar. Αυτό ισχύει περισσότερο για τον αλγόριθμο του name disambiguation, καθώς στη θέση των πραγματικών ονομάτων των συγγραφέων πολλών δημοσιεύσεων εμφανίζονται λέξεις του τίτλου, επιστημονικοί όροι ή ονόματα πανεπιστημίων.
- Αν και λόγω των citations έχουν καλυφθεί αρκετές εθνικότητες, μεγάλο μέρος των συγγραφέων εμφανίζεται μόνο μια φορά, δηλαδή σε μια μόνο δημοσίευση. Αυτό

μπορεί κάλλιστα να οδηγήσει σε πλασματικά αποτελέσματα, αφού ακόμα και ένας ακραία αυστηρός αλγόριθμος που θεωρεί όλους τους συγγραφείς διαφορετικούς μεταξύ τους επιτυγχάνει στην περίπτωση αυτή υψηλά αποτελέσματα.

- Από την άποψη του επιστημονικού κλάδου περιοριστήκαμε στην επιστήμη των υπολογιστών.

Τέλος, πρέπει να επισημάνουμε ότι ακόμα και φτωχές, μέσα σε λογικά πλαίσια φυσικά, επιδόσεις δε μειώνουν τη χρηστικότητα της εφαρμογής μας, καθώς ο χρήστης έχει τη δυνατότητα να επέμβει και να μεταβάλει τα αποτελέσματα των σχετικών αλγορίθμων κατά τη διάρκεια μιας αναζήτησης στο Web.

4

Ανάπτυξη Πληροφοριακού Συστήματος

Στο κεφάλαιο αυτό περιγράφεται συνοπτικά η διαδικασία ανάπτυξης του πληροφοριακού συστήματος, που βρίσκεται στον πυρήνα της προσπάθειας μας. Για την ακρίβεια, γίνεται μια αναφορά στην ανάλυση, τη σχεδίαση και την υλοποίηση των τμημάτων του προγράμματος που είναι επιφορτισμένα με:

- την άντληση πληροφοριών από το Google Scholar.
- την παρουσίασή τους στο χρήστη με τέτοιο τρόπο ώστε να είναι εφικτή η διαχείρισή τους.
- την καταχώρηση και ανάκτηση των επεξεργασμένων δεδομένων από τη βάση δεδομένων.

Εξαιρούμε δηλαδή τους αλγορίθμους μηχανικής μάθησης που αναπτύξαμε για την επίλυση των προβλημάτων citation matching , mixed citation και split citation και καλύψαμε στο προηγούμενο κεφάλαιο.

Άξονας της ακόλουθης ανάλυσης είναι *η προσέγγιση της σύγχρονης αντικειμενοστραφούς (object oriented) τεχνολογίας λογισμικού.* Έτσι, κάθε ενότητα είναι αφιερωμένη σε ένα από τα γενικά στάδια που αυτή καθορίζει. Πρέπει να σημειώσουμε στο σημείο αυτό ότι ίσως είναι καταχρηστικό να χρησιμοποιεί κανείς τον όρο τεχνολογία λογισμικού για ένα έργο σχετικά μικρής κλίμακας. Στόχος μας είναι όμως να δείξουμε ότι ακόμα και σε τέτοιες περιπτώσεις μια προσέγγιση βασισμένη στις αρχές της τεχνολογίας λογισμικού προσδίδει σημαντικά πλεονεκτήματα σε ένα σύστημα. Μεταξύ αυτών αναφέρουμε ενδεικτικά την απλοποιημένη δομή, την ευκολότερη συντήρηση, την επεκτασιμότητα και τη δυνατότητα επαναχρησιμοποίησης υποσυστημάτων. Με τον όρο υποσύστημα αναφερόμαστε σε πακέτα-packages, και πάλι λόγω του μικρού όγκου του έργου. Στο εξής θα χρησιμοποιούμε τον όρο πακέτο.

Πρέπει, επίσης, να επισημανθεί ότι δε χρησιμοποιούνται οι καθιερωμένοι φορμαλισμοί (UML, έγγραφα προδιαγραφών κλπ) στην προδιαγραφή των απαιτήσεων ή στα υπόλοιπα σταδία της ανάπτυξης. Αντίθετα, η περιγραφή θα στηριχθεί εξ' ολοκλήρου στη φυσική

γλώσσα και σε δομημένα κείμενα. Αυτό γίνεται αφενός λόγω έλλειψης χώρου και αφετέρου για να είναι η διαδικασία που περιγράφουμε κατανοητή και σε όσους δεν είναι εξοικειωμένοι με αυτούς τους φορμαλισμούς.

4.1 Επιλογή Μοντέλου Κύκλου Ζωής

Μια από τις κρισιμότερες παραμέτρους της ανάπτυξης ενός συστήματος αφορά στην επιλογή του κατάλληλου μοντέλου κύκλου ζωής, στο οποίο θα βασιστεί η όλη διαδικασία ανάπτυξης. Η κρισιμότητα αυτής της παραμέτρου γίνεται κατανοητή αν αντιληφθεί κανείς ότι το μοντέλο αυτό καθορίζει το σύνολο των σταδίων στα οποία υποδιαιρείται η ανάπτυξη αλλά και τις εργασίες που αυτά περιλαμβάνουν. Είναι, συνεπώς, καθοριστικός ο ρόλος του στην επιτυχή ή όχι κατάληξή του εγχειρήματος.

Στην περίπτωσή μας, επιλέχθηκε το μοντέλο πρωτοτυποποίησης, το οποίο ενδείκνυται για νέες εφαρμογές με σχετικά άγνωστες απαιτήσεις, καθώς δεν διαθέταμε προηγούμενη εμπειρία στην ανάπτυξη αντίστοιχων συστημάτων. Κεντρικό σημείο στο μοντέλο αυτό αποτελεί η δημιουργία στο πρώτο στάδιο ενός υποτυπώδους αλλά λειτουργικού πρωτότυπου συστήματος. Πράγματι, αναπτύξαμε αρχικά ένα τέτοιο πρωτότυπο, με τη βοήθεια του οποίου συλλέξαμε τις απαιτήσεις. Σημαντική ήταν επίσης και η συμβολή του στη διευκόλυνση της εργασίας των επόμενων σταδίων.

Αξίζει να επισημάνουμε ότι η σχεδίαση του συστήματος με βάση την εμπειρία του πρωτοτύπου, το κατέστησε πιο αποδοτικό, τόσο από την άποψη της ταχύτητας όσο και από την άποψη της μνήμης που καταλαμβάνει κατά τη λειτουργία του. Το γεγονός αυτό είναι σημαντικό, καθώς η απόδοση αποτελεί σημαντική παράμετρο για τη επιτυχή λειτουργία ενός συστήματος που καλείται να διαχειριστεί ένα συνεχώς αυξανόμενο όγκο δεδομένων. Επιπλέον, αν και το πλήθος των υποσυστημάτων-πακέτων αυξήθηκε από το ένα στο πέντε, το σύστημα απέκτησε τέτοια δομή ώστε να είναι συντηρήσιμο και επεκτάσιμο, χαρακτηριστικά που δεν διέθετε το αρχικό πρωτότυπο.

4.2 Προσδιορισμός των απαιτήσεων

Η υποενότητα αυτή είναι αφιερωμένη στο επόμενο στάδιο μετά την πρωτοτυποποίηση, δηλαδή των καθορισμό των απαιτήσεων του συστήματος. Οι απαιτήσεις ενός οποιουδήποτε συστήματος διακρίνονται στις εξής κατηγορίες:

- **Λειτουργικές απαιτήσεις** (*functional requirements*), που περιγράφουν τις λειτουργίες που πρέπει να επιτελεί το σύστημα.
- **Μη λειτουργικές απαιτήσεις** (*non-functional requirements*), που αποτελούν χαρακτηριστικά που πρέπει να διαθέτει το σύστημα για να θεωρείται επιτυχημένο

Το περιεχόμενο των δυο κατηγοριών, όπως αυτό διαμορφώνεται στην περίπτωσή μας, αναλύεται λεπτομερώς στις επόμενες υποενότητες.

4.2.1 Λειτουργικές απαιτήσεις

Ο καθορισμός των απαιτήσεων με τη βοήθεια του πρωτοτύπου προσδιόρισε τις ακόλουθες λειτουργίες που πρέπει να επιτελεί το σύστημα:

- On-line εύρεση όλων των δημοσιεύσεων ενός επιστήμονα και όλων των citations για κάθε μια από αυτές, με επισήμανση των νέων.
- Προβολή του συνόλου των καταχωρημένων δημοσιεύσεων ενός συγγραφέα σε συνδυασμό με τα citations τους. Οι δημοσιεύσεις να κατατάσσονται σε φθίνοντα αριθμό citations.
- Προβολή των καταχωρημένων δημοσιεύσεων ενός συγγραφέα ανά έτος σε συνδυασμό με τα αντίστοιχα citations. Οι δημοσιεύσεις κάθε έτους να κατατάσσονται σε φθίνοντα αριθμό citations.
- Προβολή όλων των συγγραφέων με τους οποίους έχει συνεργαστεί ένας συγκεκριμένος συγγραφέας σε συνδυασμό με το πλήθος των κοινών δημοσιεύσεων (co-authors). Η κατάταξη να γίνεται κατά φθίνουσα σειρά.
- Προβολή όλων των συγγραφέων που αναφέρουν τουλάχιστον μια φορά ένα συγκεκριμένο συγγραφέα στη βιβλιογραφία τους (citing authors).
- Δυναμική ανανέωση του wrapper του Google Scholar

Στους στόχους μας είναι η μελλοντική επέκταση των λειτουργιών του συστήματος με τις ακόλουθες δυνατότητες:

- Οπτικοποίηση του κοινωνικού δικτύου-γράφου που δημιουργούν οι συγγραφείς με το να αναφέρουν ο ένας εργασίες του άλλου στη βιβλιογραφία τους
- Εξαγωγή των καταχωρημένων δημοσιεύσεων ενός συγγραφέα σε συνδυασμό με τα citations τους σε αρχεία text, HTML ή XML.

4.2.2 Μη λειτουργικές απαιτήσεις

Εξίσου σημαντική με τις παραπάνω απαιτήσεις είναι και η απαίτηση για *ευχρηστία και χρηστικότητα της εφαρμογής*. Το σύστημά μας πρέπει δηλαδή να διαθέτει ένα όσο γίνεται πιο απλό και κατανοητό γραφικό περιβάλλον, ώστε η αλληλεπίδραση με το χρήστη να είναι

άμεση και αποτελεσματική χωρίς όμως να χρειάζεται προηγούμενη εξοικείωση του χρήστη με το πρόγραμμα.

Ιδιαίτερα σημαντική κρίνεται επίσης και η δυνατότητα για *εύκολη συντήρηση και επέκταση* του συστήματος με νέες δυνατότητες. Τα χαρακτηριστικά αυτά θα φροντίσουμε να διασφαλισθούν από την αρχιτεκτονική σχεδίαση που περιγράφουμε στην επόμενη υποενότητα.

Όσον αφορά στο απαραίτητο υλικό (hardware) για τον υπολογιστή στον οποίο θα τρέχει η εφαρμογή, δεν υπάρχει κάποια ιδιαίτερη απαίτηση λόγω του μικρού όγκου της εφαρμογής. Αυτονόητη κρίνεται ωστόσο η δυνατότητα σύνδεσης στο Διαδίκτυο.

Τέλος, στις επόμενες γραμμές περιγράφουμε τις τεχνολογίες που χρησιμοποιήθηκαν στην ανάπτυξη του συστήματος.

Η γλώσσα προγραμματισμού που επιλέξαμε για τη δημιουργία της εφαρμογής είναι η **Java (έκδοση 1.5)**, λόγω των ακόλουθων πλεονεκτημάτων της:

- Είναι συμβατή με όλα τα λειτουργικά συστήματα (*μεταφερσιμότητα-portability*).
- Χαίρει ευρείας αποδοχής από την επιστημονική κοινότητα. Υπάρχει μάλιστα πληθώρα εξαιρετικών βιβλιοθηκών ανοικτού κώδικα (open source) που υλοποιούν συναρτήσεις μηχανικής μάθησης (π.χ. SecondString, Weka, libSVM) ή της προσθέτουν επιπλέον δυνατότητες (π.χ. σχεδιασμός γράφων μέσω της JUNG). Αυτό το στοιχείο προσθέτει επομένως μεγάλες δυνατότητες επέκτασης και βελτίωσης σε ένα σύστημα γραμμένο σε Java, ενώ επιταχύνει σημαντικά και την ανάπτυξή του.
- Προσφέρει εξαιρετικά απλούς και χρηστικούς τρόπους για την αντιμετώπιση πολύ διαφορετικών προβλημάτων, όπως γραφικά, επεξεργασία αρχείων κάθε είδους, parsing ιστοσελίδων κλπ.

Το περιβάλλον ανάπτυξης που χρησιμοποιήσαμε είναι το **NetBeans (έκδοση 5.0)**, το οποίο διακρίνεται για τη φιλική διεπαφή του και τη σταθερότητά του.

Για την ανάπτυξη και διαχείριση της βάσης δεδομένων χρησιμοποιήσαμε τη **MySQL (έκδοση 5.0)**, επειδή είναι open source και διαθέτει πληθώρα συναρτήσεων για το χειρισμό των διαφόρων τύπων δεδομένων. Για την ακρίβεια, χρησιμοποιήσαμε τα βοηθητικά εργαλεία **MySQL GUI Tools** (MySQL Administrator 1.2.4 και MySQL Query Browser 1.2.4beta).

4.3 Αρχιτεκτονική Σχεδίαση (Architectural Design)

Η ενότητα αυτή είναι αφιερωμένη στην αρχιτεκτονική σχεδίαση του συστήματός μας. Θα περιγραφεί δηλαδή λεπτομερώς η γενική δομή της εφαρμογής (τα υποσυστήματα-πακέτα από τα οποία αποτελείται), όπως αυτή διαμορφώθηκε για να αντεπεξέλθει και να καλύψει το

σύνολο των παραπάνω (λειτουργικών και μη) απαιτήσεων. Η ανάλυσή μας δεν υπεισέρχεται ωστόσο σε λεπτομέρειες της σχεδίασης, όπως η εσωτερική δομή των κλάσεων.

Βασικοί άξονες της σχεδίασης αυτής είναι η υψηλή συνεκτικότητα των κλάσεων κάθε πακέτου και η χαμηλή σύζευξη μεταξύ των πακέτων. Τα δύο αυτά χαρακτηριστικά εξασφαλίζουν αντίστοιχα ότι κάθε πακέτο σχετίζεται με μια συγκεκριμένη λειτουργία και ότι δεν υπάρχει μεγάλη αλληλεξάρτηση μεταξύ των διαφορετικών πακέτων. Πιο συγκεκριμένα, το σύστημά μας χαρακτηρίζεται από **λειτουργική συνεκτικότητα**, η οποία αφενός καθιστά ορισμένα πακέτα επαναχρησιμοποιήσιμα και αφετέρου διευκολύνει τον εντοπισμό λαθών, και **σύζευξη δεδομένων**, ώστε να μεταφέρονται μέσω των διεπαφών μόνο τα απαραίτητα δεδομένα. Λαμβάνεται επιπλέον ειδική μέριμνα ώστε οι διεπαφές, δηλαδή ο τρόπος κλήσης μεθόδων και υπηρεσιών, να είναι απλές και σαφείς.

Στη συνέχεια, παρουσιάζουμε σε ξεχωριστές υποενότητες τα υποσυστήματα στα οποία υποδιαιρέσαμε την εφαρμογή μας με βάση τα παραπάνω κριτήρια. Δεν αναφερόμαστε ωστόσο σε συγκεκριμένες αρχιτεκτονικές τεχνοτροπίες, καθώς ουσιαστικά προκύπτει ένας περίπλοκος συνδυασμός ορισμένων από αυτών (σημειακά ετερογενή αρχιτεκτονική). Αναφέρουμε ενδεικτικά κάποιες που χρησιμοποιήσαμε : call & return, interacting processes–threads & Model View Controller, data oriented repository-transactional databases.

4.3.1 Γραφικό Περιβάλλον (GUI Package)

Όπως υποδηλώνει και το όνομά του, το πακέτο αυτό περιέχει τις κλάσεις που είναι επιφορτισμένες με την παρουσίαση των δεδομένων στο χρήστη αλλά και την αλληλεπίδραση μαζί του. Προσφέρει δηλαδή τις απαραίτητες επιλογές για την ικανοποίηση των λειτουργικών απαιτήσεων του συστήματος.

Στη συνέχεια απαριθμούμε τις κλάσεις που περιέχει αυτό το πακέτο, σχολιάζοντας συνοπτικά τη λειτουργία τους:

- ***MainFrame***: δημιουργεί και εμφανίζει το παράθυρο της εφαρμογής, το οποίο περιέχει όλες τις διαθέσιμες στο χρήστη επιλογές είτε μέσω ενός μενού είτε μέσω tabs. Περιέχει επίσης τη μέθοδο main που θέτει σε λειτουργία το σύστημα.
- ***OnLineAbstractTab, OnLineCitationSearch, OnLineSearchTab***: η πρώτη κλάση είναι ο abstract πρόγονος των υπόλοιπων που καθορίζει τις κοινές τους μεθόδους. Οι δυο τελευταίες κλάσεις δημιουργούν τα tab που παρέχουν τη δυνατότητα στο χρήστη να απευθύνει ένα ερώτημα στο Google Scholar για τις δημοσιεύσεις και τα citations αντίστοιχα ενός επιστήμονα. Δίνουν επίσης τη δυνατότητα στο χρήστη να επεξεργαστεί και να επιβεβαιώσει τα αποτελέσματα των αλγορίθμων επίλυσης των προβλημάτων citation matching, mixed citation και split citation.

- ***ProcessStoredDataTab***: δημιουργεί το tab που δίνει τη δυνατότητα στο χρήστη να δει στοιχεία για τους επιστήμονες που είναι καταχωρημένοι στη βάση δεδομένων, όπως δημοσιεύσεις, συνώνυμα, co-authors και citing authors.
- ***UpdateGSWrapperTab***: εμφανίζει το tab, μέσω του οποίου γίνεται η ενημέρωση του wrapper του Google Scholar (περισσότερα στην υποενότητα 4.3.4)
- ***CitationMatchingFrame***: το παράθυρο αυτό περιέχει τα ζευγάρια δημοσιεύσεων που ο αντίστοιχος αλγόριθμος έχει υποδείξει ότι κατά πάσα πιθανότητα ταυτίζονται, παρότι εμφανίζονται με διαφορετικό το τίτλο. Παρέχει έτσι στο χρήστη τη δυνατότητα να διορθώσει τα αποτελέσματα του αλγορίθμου αποεπιλέγοντας αυτά που κατά τη γνώμη του δεν αφορούν την ίδια δημοσίευση.
- ***CitationsFrame***: δημιουργεί και εμφανίζει το παράθυρο που παρουσιάζει τα citations μιας συγκεκριμένης εργασίας με βάση στοιχεία της ΒΔ. Παρέχει επίσης τη δυνατότητα ταξινόμησης των citations ανά έτος δημοσίευσης ή/και δικό τους αριθμό citation.
- ***SynonymsFrame***: δημιουργεί το παράθυρο που παρουσιάζει τα συνώνυμα του δοσμένου συγγραφέα που είτε βρίσκονται καταχωρημένα στη ΒΔ είτε προκύπτουν από μια on-line αναζήτηση.
- ***PaperPanel***: δημιουργεί ένα panel που εμφανίζει με δομημένο τρόπο τα πεδία μιας δοσμένης δημοσίευσης.
- ***LightPaperPanel***: αποτελεί μια παραλλαγή της προηγούμενης κλάσης, που δημιουργεί ένα λιγότερο διαδραστικό panel για τη δοσμένη δημοσίευση.
- ***MergeOrNotPanel***: δημιουργεί ένα απλό panel με δυο RadioButtons για να διαλέξει ο χρήστης αν θα "συγχωνεύσει" ή όχι δυο δημοσιεύσεις.
- ***PaperToBeMergedPanel***: εμφανίζει τα απαραίτητα στοιχεία (τίτλο, URLs και citation URLs) για μπορέσει ο χρήστης να ταυτοποιήσει μια υπό συγχώνευση δημοσίευση.
- ***MergerPanel***: χρησιμοποιεί τις δυο προηγούμενες κλάσεις για να επιτρέψει στο χρήστη να αποφασίσει αν μια προτεινόμενη "συγχώνευση" δημοσιεύσεων είναι ορθή. Χρησιμοποιείται από το CitationMatchingFrame.
- ***ScholarPanel***: εμφανίζει τα στοιχεία (όνομα, συνώνυμα, δημοσιεύσεις) για ένα επιστήμονα όπως αυτά προκύπτουν μετά από την επίλυση των mixed και split citation προβλημάτων από τον αντίστοιχο αλγόριθμο. Παρέχει ταυτόχρονα τη δυνατότητα στο χρήστη να διορθώσει τα αποτελέσματά τους.
- ***ScholarSideBar***: αποτελείται από αντικείμενα της προηγούμενης κλάσης (ένα για κάθε διαφορετικό επιστήμονα) και εμφανίζεται μετά την επίλυση των MCP και SCP
- ***ScholarTab***: εμφανίζει τις δημοσιεύσεις ενός επιστήμονα, ομαδοποιημένες ανά συνώνυμο, όπως αυτές διαμορφώνονται μετά την επίλυση των προβλημάτων mixed και split citation.

- ***CompletableJTextField*** : δημιουργεί το textfield, στο οποίο όταν εισάγονται γράμματα, εμφανίζεται ένα παραθυράκι με τους επιστήμονες που το όνομά τους ξεκινά με τα δοσμένα γράμματα ([MA05]).
- ***StatusBar***: δημιουργεί τη μπάρα που εμφανίζεται στο κάτω μέρος της οθόνης και ενημερώνει το χρήστη για την πορεία της on-line αναζήτησης ([MA05]).
- ***Constants***: είναι το interface που περιέχει σε μορφή σταθερών διάφορα βοηθητικά στοιχεία για το γραφικό περιβάλλον.
- ***Utilities***: η κλάση αυτή περιέχει στατικές μεθόδους γενικής χρήσης.
- ***VectorButton, RedButton, BlueButton***: κλάσεις για τη δημιουργία κουμπιών. Οι δύο τελευταίες είναι απόγονοι της πρώτης ([MA05]).
- ***CitationButton, CitationMatchingButton***: οι κλάσεις αυτές είναι και οι δύο απόγονοι της BlueButton. Ο ρόλος τους είναι να εμφανίζουν ένα CitationFrame και ένα παράθυρο με τους διαφορετικούς τίτλους κάθε δημοσίευσης αντίστοιχα.
- ***URLComboBox****:* δημιουργεί ένα combobox με την ιδιότητα ότι όταν ένα στοιχείο του (URL) επιλεγεί, ανοίγει ο default browser στο επιλεγμένο URL.
- ***EnhancedComboBox***: δημιουργεί combobox με στοιχεία τίτλους δημοσιεύσεων. Όταν ένας τίτλος επιλεγεί από το χρήστη, προβάλει ένα CitationFrame για την αντίστοιχη δημοσίευση .

4.3.2 Επικοινωνία με Βάση Δεδομένων (DBMS Package)

Το συγκεκριμένο υποσύστημα-πακέτο περιέχει τις κλάσεις που αναλαμβάνουν να διεκπεραιώσουν την επικοινωνία της εφαρμογής με τη Βάση Δεδομένων, καταχωρώντας και ανακτώντας τα επεξεργασμένα δεδομένα που προκύπτουν από τις on-line αναζητήσεις. Η διάρθρωση του πακέτου σε κλάσεις στηρίχθηκε στο πεπερασμένο σύνολο ειδών ενεργειών που μπορούν να εκτελεστούν σε μια βάση δεδομένων (εισαγωγή, διαγραφή, ενημέρωση, εύρεση) και διαμορφώνεται ως εξής:

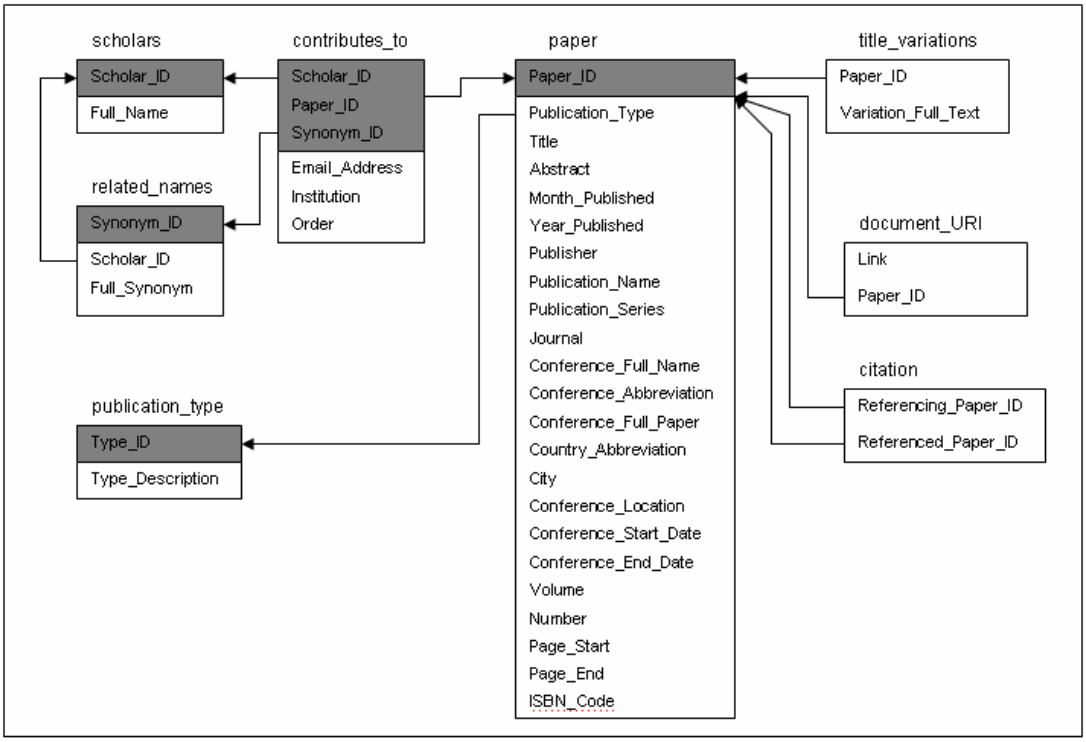
- ***DatabaseManager***: μοναδικό μέλημα της κλάσης αυτής είναι η σύνδεση και η αποσύνδεση με την ΒΔ.
- ***DeleteStatements***: περιέχει τις εντολές διαγραφής δεδομένων από τη ΒΔ.
- ***InsertStatements***: περιέχει μια εντολή εισαγωγής για κάθε πίνακα της ΒΔ.
- ***SelectStatements*** : περιέχει το σύνολο των εντολών με τις οποίες αντλούνται δεδομένα τη ΒΔ.
- ***UpdateStatements***: περιλαμβάνει τις εντολές ενημέρωσης δεδομένων.
- ***SophisticatedStatements***: οι μέθοδοι της κλάσης αυτής εκτελούν τις σύνθετες λειτουργίες της εφαρμογής που σχετίζονται με τη ΒΔ, όπως αποθήκευση όλων των

δεδομένων που αφορούν μια δημοσίευση ή ένα επιστήμονα. Για τη λειτουργία τους συνδυάζονται οι στοιχειώδεις εντολές των προηγούμενων κλάσεων.

- ***Database***: η κλάση αυτή περιέχει ως πεδία της ένα αντικείμενο για κάθε μια από τις προηγούμενες κλάσεις και σκοπός της είναι να συνδυάσει όλες τις λειτουργίες τους σε μια και μόνο κλάση. Πιο συγκεκριμένα, η εμβέλεια των μεθόδων των παραπάνω κλάσεων περιορίζεται στο πακέτο DBMS αλλά μέσω της συγκεκριμένης κλάσης διατίθεται σε όλο το σύστημα.

Ιδιαίτερη μέριμνα έχει ληφθεί σε όλες τις μεθόδους που τροποποιούν τα δεδομένα της ΒΔ για την αποφυγή διπλοεγγραφών (*duplicate records*).

Στο σημείο αυτό πρέπει να παραθέσουμε το διάγραμμα σχήματος της ΒΔ (ελλείψει καθιερωμένου προτύπου θα χρησιμοποιήσουμε τον τρόπο που προτείνεται στο [SKS02]). Από το διάγραμμα αυτό κρίθηκε σκόπιμο να παραληφθούν οι όψεις που έχουν δημιουργηθεί (για βοηθητικούς και μόνο λόγους) και ο πίνακας στον οποίο καταχωρούνται οι απαραίτητες πληροφορίες για τον wrapper.



Εικόνα 4. Διάγραμμα Σχήματος της ΒΔ

Παρακάτω επιχειρείται μια σύντομη ερμηνεία των πινάκων του παραπάνω σχήματος.

- **paper**: διαθέτει πλήθος ιδιοτήτων, οι οποίες καλύπτουν όλο το φάσμα των πληροφοριών που μπορεί να βρει κανείς για μια δημοσίευση. Μόνο ένα μικρό μέρος από αυτές συμπληρώνονται ακόμα και στην καλύτερη περίπτωση από τα δεδομένα του Google Scholar. Διατηρήθηκαν ωστόσο με το σκεπτικό ότι ίσως μελλοντικά προστεθεί

η δυνατότητα να τις ορίζει ο χρήστης. Επιπλέον, ορισμένες από αυτές ενδεχομένως να συμπληρώνονται στο μέλλον από δεδομένα άλλων ακαδημαϊκών μηχανών αναζήτησης.

- **title_variations**: περιέχει τους διαφορετικούς τίτλους που μπορεί να έχει μια δημοσίευση εξαιτίας του Citation Matching προβλήματος.
- **document_uri**: στον πίνακα αυτόν καταχωρούνται τα URL στα οποία είναι διαθέσιμη μια δημοσίευση.
- **citation**: ο πίνακας αυτός μοντελοποιεί την βιβλιογραφική αναφορά που περιέχει μια εργασία (referencing paper) σε μια άλλη (referenced paper).
- **scholar**: οι εγγραφές του πίνακα αυτού αντιστοιχούν στους διαφορετικούς επιστήμονες που έχουν εντοπιστεί μετά την επίλυση των Mixed και Split Citation προβλημάτων. Έχει επιλεγεί σε κάθε ένα επιστήμονα να αντιστοιχίζεται το μεγαλύτερο από τα διαφορετικά ονόματα με τα οποία εμφανίζεται, ως πιο αντιπροσωπευτικό.
- **related_names**: στον πίνακα αυτό καταχωρούνται τα διαφορετικά ονόματα με τα οποία συναντάται ο ίδιος επιστήμονας στη βιβλιογραφία, λόγω του Split Citation προβλήματος.
- **contributes_to**: ο πίνακας αυτός μοντελοποιεί τη συνεισφορά ενός επιστήμονα σε μια δημοσίευση. Περιέχει επομένως τα στοιχεία που αναγράφονται σε μια δημοσίευση για κάθε συγγραφέα, όπως ο λογαριασμός του e-mail του και το ίδρυμα στο οποίο ανήκει. Καταχωρείται επίσης και η σειρά με την οποία εμφανίζεται το όνομά του. Βέβαια η συμπλήρωση όλων αυτών των λεπτομερειών από τα στοιχεία που αντλούμε από τον Παγκόσμιο Ιστό δεν είναι εφικτή στην πράξη. Και πάλι όμως δεν αποκλείεται μελλοντικά η προσθήκη της δυνατότητας καθορισμού τους από το χρήστη.

4.3.3 Διαχείριση πληροφοριών (InfoManagement Package)

Όπως υποδηλώνει και το όνομά του, σε αυτό το πακέτο έχουν περιληφθεί οι κλάσεις που σχετίζονται με την επεξεργασία των δεδομένων που αντλούνται από τον Παγκόσμιο Ιστό ή τη Βάση Δεδομένων. Πιο συγκεκριμένα, οι κλάσεις αυτές είναι οι εξής :

- ***Paper***: η κλάση αυτή έχει ως ιδιότητες τα πεδία μιας δημοσίευσης όπως προκύπτουν είτε από μια αναζήτηση στο Web είτε από τη ΒΔ. Διαθέτει επίσης στοιχειώδεις μεθόδους για το χειρισμό αυτών των ιδιοτήτων (setX, getX). Στο σημείο αυτό είναι απαραίτητο να επισημάνουμε ποιες ακριβώς πληροφορίες μπορούμε να αντλήσουμε *στην βέλτιστη περίπτωση* από το Google Scholar. Αναφερόμαστε στη βέλτιστη περίπτωση γιατί, όπως θα καταστεί σαφές από την επόμενη ενότητα, δεν περιέχουν όλες οι εγγραφές του Google Scholar τις ίδιες πληροφορίες:
  - Τίτλος
  - Συγγραφείς

- Abstarct
- URLs

- Έτος και Μήνας Δημοσίευσης
- Τύπος Δημοσίευσης (βλέπε κατηγοριοποίηση στην ενότητα 1.1)
- Τίτλος συνεδρίου ή περιοδικού όπου δημοσιεύθηκε
- Τόμος (volume),
- Έκδοση (issue)
- Σελίδες περιοδικού ή πρακτικών
- Εκδότης

Όπως φαίνεται, έχουμε χωρίσει τα πεδία σε δύο κατηγορίες. Η πρώτη περιλαμβάνει τα πεδία που αφενός είναι απαραίτητα για την μετέπειτα επεξεργασία των δεδομένων (επίλυση των προβλημάτων Citation Matching, Mixed και Split Citation) και αφετέρου είναι αυτά που οι περισσότερες εγγραφές του Google Scholar περιέχουν. Αντίθετα, τα πεδία της δεύτερης ομάδας είναι πιο επισφαλή, δηλαδή σε σπάνιες περιπτώσεις καταφέρνουμε να τα συμπληρώσουμε. Για τα δυο πρώτα από αυτά πάντως (έτος και τύπος δημοσίευσης) καταβάλλεται ιδιαίτερη προσπάθεια μέσω ειδικών συναρτήσεων.

- ***PapersOfYear***: ο ρόλος που έχει η συγκεκριμένη κλάση είναι να εντοπίζει στη ΒΔ τις δημοσιεύσεις ενός συγκεκριμένου συγγραφέα ή τα citations συγκεκριμένου paper που εκδόθηκαν στη διάρκεια ενός συγκεκριμένου έτους**.** Αν δεν παρέχεται συγκεκριμένο έτος, βρίσκει το σύνολο των σχετικών δημοσιεύσεων, ανεξαρτήτως έτους δημοσίευσης.
- ***PapersOfAllYears***: σκοπός της κλάσης αυτής είναι να βρει από τη ΒΔ όλα τα έτη κατά οποία δημοσίευσε εργασίες ο δοσμένος συγγραφέας ή έγιναν citations σε ένα paper. Στη συνέχεια εντοπίζει για κάθε ένα από τα έτη αυτά τις αντίστοιχες εργασίες χρησιμοποιώντας την προηγούμενη κλάση και τις εμφανίζει στο ProcessStoredDataTab.
- ***Author***: η κλάση αυτή περιέχει τα απαραίτητα πεδία για κάθε ένα από τους διαφορετικούς επιστήμονες που προκύπτουν από τη λύση των mixed και slit citation προβλημάτων και έχουν καταχωρηθεί στη ΒΔ. Τέτοια πεδία είναι τα συνώνυμά του, το πιο αντιπροσωπευτικό όνομα και οι δημοσιεύσεις στις οποίες συμμετέχει.
- ***PaperContribution***: μέσω αυτής της κλάσης μοντελοποιείται με τα κατάλληλα πεδία, η συμμετοχή ενός συγγραφέα σε μια δημοσίευση. Χρησιμοποιείται από την προηγούμενη κλάση.
- ***CoAuthors***: η κλάση αυτή αναλαμβάνει να εντοπίσει στη ΒΔ όλους τους επιστήμονες με τους οποίους έχει συνεργαστεί ο δοσμένος επιστήμονας και να τους ταξινομήσει με βάση τον αριθμό κοινών δημοσιεύσεων. Τα αποτελέσματα εμφανίζονται στο ProcessStoredDataTab.

- ***CitingAuthors***: η κλάση αυτή είναι υπεύθυνη για την εύρεση των επιστημόνων που έχουν αναφέρει τουλάχιστον μια φορά στη βιβλιογραφία τους μια δημοσίευση του δοσμένου επιστήμονα. Τους επιστήμονες αυτούς τους ταξινομεί σε φθίνουσα σειρά citations και στη συνέχεια τους εμφανίζει στο ProcessStoredDataTab.
- ***CorrelatedAuthor***: η κλάση αυτή περιέχει τα κοινά πεδία των co-authors ή citing authors. Χρησιμοποιείται από τις αντίστοιχες προηγούμενες κλάσεις.
- ***Scholar***: η κλάση αυτή χρησιμοποιείται κατά την επίλυση των Mixed και Split Citation προβλημάτων. Πιο συγκεκριμένα, δέχεται μια συνεκτική συνιστώσα (connected component) του γράφου συγγραφέων που σχηματίζεται από τον αρμόδιο αλγόριθμο μηχανικής μάθησης και *δημιουργεί ένα ξεχωριστό επιστήμονα*. Βρίσκει, δηλαδή, το πιο αντιπροσωπευτικό του όνομα, τα διαφορετικά συνώνυμά του και τις δημοσιεύσεις που αντιστοιχούν σε κάθε συνώνυμο. Διαθέτει επίσης τις απαραίτητες συναρτήσεις για επεξεργασία των αποτελεσμάτων του αλγορίθμου από το χρήστη.
- ***Constants***: το interface αυτό, όπως υποδηλώνει το όνομά του, περιλαμβάνει μια σειρά σταθερών, που αφορούν στα δεδομένα των προηγούμενων κλάσεων, (π.χ. default τιμές για κάποια πεδία της Paper.java).

4.3.4 Wrapper του Google Scholar (GSWrapper Package)

Το πακέτο αυτό περιλαμβάνει το σύνολο των κλάσεων που σχετίζονται με την εξαγωγή δεδομένων από το Google Scholar αλλά και την ενημέρωση του wrapper. Πριν όμως παρουσιάσουμε τις κλάσεις του πακέτου, πρέπει στο σημείο αυτό να διευκρινίσουμε τα ακόλουθα σημεία για να καταστήσουμε σαφή το λόγο ύπαρξής τους:

- **Είδη των δημοσιεύσεων που επιστρέφει το Google Scholar**
  Οι εγγραφές που επιστρέφει το Google Scholar δεν χαρακτηρίζονται από ομοιογένεια. Αντίθετα, επειδή, όπως ήδη αναφέραμε, πρόκειται για μια "ανώριμη" υπηρεσία, ποικίλουν αρκετά στο περιεχόμενο τους αλλά και στην ποιότητα των πληροφοριών τους. Για την ακρίβεια εντοπίσαμε τέσσερις κατηγορίες με βασικά κριτήρια το *πλήθος των URL τους* και την *ύπαρξη ή όχι abstract*. Επιλέξαμε τα συγκεκριμένα κριτήρια, επειδή αυτά καθορίζουν σε μεγάλο βαθμό τον τρόπο εξαγωγής πληροφοριών, ενώ συγχρόνως διαδραματίζουν σημαντικό ρόλο και στην επεξεργασία των δημοσιεύσεων (λύση citation matching).

- ***Κατηγορία 1:*** *πλήθος URL > 1, ύπαρξη abstract*


An evaluation of Naive Bayesian anti-spam filtering - all 15 versions »
... , J Koutsias, KV Chandrinos, **G Paliouras**, CD ... - Arxiv preprint cs.CL/0006013, 2000 - arxiv.org
Proceedings of the workshop on Machine Learning in the New Information Age, G.
Potamias, V. Moustakis and M. van Someren (eds.), 11 th European Conference on
Machine Learning, Barcelona, Spain, pp. 9-17, 2000.
Cited by 146 - Related Articles - View as HTML - Web Search - Import into BibTeX - Library Search

Εικόνα 5. Παράδειγμα εγγραφής Google Scholar κατηγορίας 1

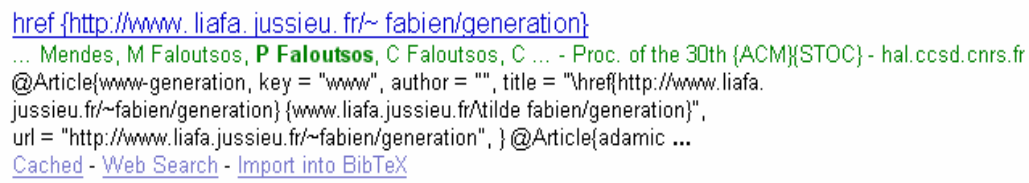
Όπως διαπιστώνει κανείς από το παραπάνω παράδειγμα, οι εγγραφές αυτού του τύπου χαρακτηρίζονται από υψηλή ποιότητα και ποσότητα πληροφοριών, δηλαδή περιέχουν συνήθως σωστές πληροφορίες και μάλιστα για το σύνολο σχεδόν των πεδίων μιας δημοσίευσης.

- ***Κατηγορία 2:*** *πλήθος URL = 1, ύπαρξη abstract*


E-grids: Computationally efficient grammatical inference from positive examples
G Petasis, **G Paliouras**, V Karkaletsis, C Halatsis, ... - Grammars, 2004 - grammars.grlmc.com
Georgios Petasis Georgios Paliouras Vangelis Karkaletsis Software and Knowledge
Engineering Laboratory, Institute of Informatics and Telecommunications,
National Centre for Scientific Research (NCSR) "Demokritos", PO BOX ...
Cited by 8 - Related Articles - View as HTML - Web Search - Import into BibTeX - Library Search

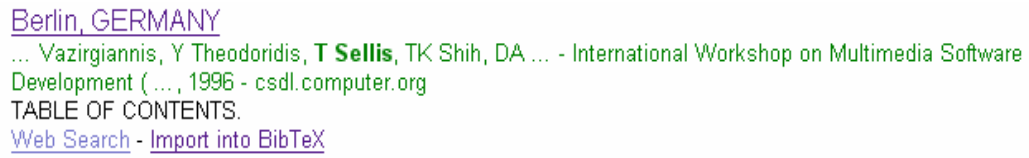
Εικόνα 6. Παράδειγμα εγγραφής Google Scholar κατηγορίας 2

Η κύρια διαφορά της από την προηγούμενη κατηγορία έγκειται στον τρόπο εξαγωγής του URL της δημοσίευσης. Είναι επίσης υψηλότερη η πιθανότητα εσφαλμένων πληροφοριών, όπως συμβαίνει στην ακόλουθη περίπτωση :


href {http://www. liafa. jussieu. fr/~ fabien/generation}
... Mendes, M Faloutsos, **P Faloutsos**, C Faloutsos, C ... - Proc. of the 30th {ACM}{STOC} - hal.ccsd.cnrs.fr
@Article{www-generation, key = "www", author = "", title = "\href{http://www.liafa.
jussieu.fr/~fabien/generation} {www.liafa.jussieu.fr/\tilde fabien/generation}",
url = "http://www.liafa.jussieu.fr/~fabien/generation", } @Article{adamic ...
Cached - Web Search - Import into BibTeX

Εικόνα 7. Παράδειγμα εγγραφής Google Scholar κατηγορίας 2

- ***Κατηγορία 3:*** *πλήθος URL = 1, μικρό abstract (λιγότερες από 10 λέξεις)*


Berlin, GERMANY
... Vazirgiannis, Y Theodoridis, **T Sellis**, TK Shih, DA ... - International Workshop on Multimedia Software
Development (... , 1996 - csdl.computer.org
TABLE OF CONTENTS.
Web Search - Import into BibTeX

Εικόνα 8. Παράδειγμα εγγραφής Google Scholar κατηγορίας 3

Οι πληροφορίες που διαθέτουν οι δημοσιεύσεις της κατηγορίας αυτής περιορίζονται συνήθως στα βασικά πεδία (συγγραφείς, τίτλος, URL), ενώ είναι ιδιαίτερα υψηλή η πιθανότητα να είναι λανθασμένες. Τα συχνότερα λάθη είναι να εμφανίζονται σε διαφορετικά από τα σωστά πεδία οι πληροφορίες (π.χ. στη θέση του τίτλου το όνομα ενός συγγραφέα ή ο τόπος διεξαγωγής ενός συνεδρίου).

- ***Κατηγορία 4:*** *πλήθος URL = 0, μικρό abstract (λιγότερες από 10 λέξεις)*

Εικόνα 9. Παράδειγμα εγγραφής Google Scholar κατηγορίας 4

Η συντριπτική πλειοψηφία της κατηγορίας αυτής (χαρακτηρίζεται [CITATION] από το Google Scholar) περιέχει λανθασμένες πληροφορίες. Η συχνότερη περίπτωση αφορά τίτλους δημοσιεύσεων με επισυναπτόμενο στο τέλος ή την αρχή τους τον τίτλο του συνεδρίου ή όνομα ενός συγγραφέα. Συχνό είναι επίσης το φαινόμενο ο τίτλος της δημοσίευσης να εμφανίζεται στη θέση του τίτλου του περιοδικού/συνεδρίου.

- **Τρόπος επεξεργασίας των δημοσιεύσεων ανάλογα με το είδος τους**

Σε γενικές γραμμές, η επεξεργασία μιας HTML σελίδας αποτελεσμάτων του Google Scholar έχει ως εξής : η αρχική σελίδα διαχωρίζεται σε υποσυμβολοσειρές (substrings), κάθε μια από τις οποίες αντιστοιχεί σε μια δημοσίευση. Οι υποσυμβολοσειρές αυτές αναλύονται διαδοχικά από μια συνάρτηση, η οποία αναλαμβάνει να εντοπίσει την διαθέσιμη πληροφορία για κάθε πεδίο της δημοσίευσης. Η ανάκτηση πληροφορίας για κάθε ένα από τα πεδία αυτά εξαρτάται, όπως αναφέραμε και παραπάνω, από την κατηγορία στην οποία ανήκουν. Συνοπτικά, γίνεται ως εξης:

Ο τρόπος ανάκτησης του *τίτλου της δημοσίευσης* είναι κοινός για όλες τις κατηγορίες δημοσιεύσεων: εξάγεται από την HTML σελίδα με τα αποτελέσματα που επιστρέφει το Google Scholar. Από την ίδια HTML σελίδα ανακτάμε και το *abstract για τις δημοσιεύσεις* των κατηγοριών 1 και 2.

Ομοίως, κοινός είναι ο τρόπος εξαγωγής των *υπόλοιπων πεδίων εκτός των URLs*. Αναλυτικότερα, οι πληροφορίες αυτές εξάγονται από την **εγγραφή BibTex** της αντίστοιχης δημοσίευσης. Τέτοια εγγραφή διαθέτουν όλες οι δημοσιεύσεις ανεξαρτήτως κατηγορίας, εφόσον βέβαια έχει ενεργοποιηθεί σχετική επιλογή του Google Scholar (χρειάζεται δηλαδή το κατάλληλο cookie). Οι εγγραφές αυτές είναι στην ουσία απλό κείμενο (text), με καθορισμένη όμως μορφή και πεδία. Ενδεικτική της μορφής τους είναι η ακόλουθη εγγραφή, η οποία τυχαίνει να διαθέτει πληροφορίες για όλα τα πεδία:

```
@article{sakkis2003mba,
  title={{A Memory-Based Approach to Anti-Spam Filtering for Mailing Lists}},
  author={Sakkis, G. and Androutsopoulos, I. and Paliouras, G. and Karkaletsis, V. and
Spyropoulos, C.D. and Stamatopoulos, P.},
  journal={Information Retrieval},
  volume={6},
  number={1},
  pages={49--73},
  year={2003},
  publisher={Springer}
}
```

Τέλος, σημαντικά διαφοροποιείται η *εξαγωγή των URL* μιας δημοσίευσης της κατηγορίας 1 από την αντίστοιχη διαδικασία για τις δημοσιεύσεις των κατηγοριών 2 και 3. Πιο συγκεκριμένα, το URL των δημοσιεύσεων των κατηγοριών 2 και 3 εξάγεται από την αρχική HTML σελίδα, όπως και ο τίτλος και το abstract. Αντίθετα, η πρώτη κατηγορία απαιτεί την επεξεργασία μιας νέας HTML ιστοσελίδας, με εγγραφές σαν το επόμενο παράδειγμα.


An Evaluation of Naive Bayesian Anti-Spam Filtering
I Androutsopoulos, J Koutsias, KV Chandrinos, G ... - iit.demokritos.gr
It has recently been argued that a Naive Bayesian classifier can be used to filter unsolicited bulk e-mail ("spam"). We conduct a thorough evaluation of this proposal on a corpus that we make publicly available, contributing ...
View as HTML - Web Search - Import into BibTeX

Εικόνα 10. Παράδειγμα εγγραφής για εξαγωγή του URL μιας δημοσίευσης

- **Τρόπος ανάπτυξης και συντήρησης του wrapper** (*wrapper maintance*)

  Αρχικά, για την *ανάπτυξη του wrapper* του Google Scholar επιδιώξαμε να χρησιμοποιήσουμε κάποιο από τα διαθέσιμα εργαλεία αυτόματης ή ημιαυτόματης δημιουργίας. Τα αποτελέσματα όλων όμως ήταν φτωχά, παρ' όλο που υπάρχει ικανοποιητική κανονικότητα στα tags των επιστρεφόμενων HTML σελίδων. Έτσι, τελικά επιλέξαμε να αναπτύξουμε το wrapper χειρωνακτικά και μάλιστα χωρίς τη χρήση κανονικών εκφράσεων. Αντίθετα, αξιοποιήθηκαν οι συναρτήσεις χειρισμού συμβολοσειρών που διαθέτει η Java.

  Όσον αφορά την *συντήρηση του wrapper*, επιλέξαμε να γίνεται και αυτή χειρωνακτικά, αφού άλλωστε αφορά μια μόνο ιστοσελίδα. Πιο συγκεκριμένα, η δυσλειτουργία του προγράμματος γίνεται άμεσα αντιληπτή όταν το πρόγραμμα αδυνατεί να εξάγει πληροφορία για κάποια πεδία και εγείρονται τα αντίστοιχα exceptions της Java. Στην περίπτωση αυτή προτρέπεται με σχετικό μήνυμα ο χρήστης να ενημερώσει το wrapper. Το παράθυρο που έχει δημιουργηθεί για το σκοπό αυτό, UpdateGSWrapperTab, εμφανίζει και επεξηγεί τις συμβολοσειρές στις οποίες βασίζεται η λειτουργία του wrapper. Συνεπώς, ο χρήστης είναι σε θέση να εντοπίσει τη νέα μορφή των συμβολοσειρών αυτών εξετάζοντας τον πηγαίο HTML κώδικα μιας σελίδας

αποτελεσμάτων του Google Scholar. Οι αλλαγές που κάνει ο χρήστης ενημερώνουν άμεσα τον wrapper, συνεπώς δε χρειάζεται επανεκκίνηση της εφαρμογής για να ελέγξει ο χρήστης την ορθότητα των αλλαγών του. Πρέπει, επίσης, να τονίσουμε στο σημείο αυτό ότι δίνεται στο χρήστη η δυνατότητα μεταβολής των τιμών όλων των σημαντικών για τη λειτουργία του wrapper πεδίων, των χαρακτηριστικών δηλαδή tags της HTML που οριοθετούν τα δεδομένα. Οι τιμές των πεδίων αυτών καταχωρούνται σε ένα στοιχειώδη πίνακα της ΒΔ, καθιστώντας έτσι εύκολη την ενημέρωσή τους.

- **Περιορισμοί του Google Scholar**

  Για διάφορους λόγους, όπως συμφωνίες με τους προμηθευτές του περιεχομένου, η Google επιβάλει δύο περιορισμούς στη χρήση της συγκεκριμένης υπηρεσίας της:

  ❖ Για κάθε αναζήτηση, είτε αυτή αφορά αναζήτηση δημοσιεύσεων ενός επιστήμονα είτε αναζήτηση των citations μιας δημοσίευσης, επιστρέφονται το πολύ 1000 αποτελέσματα. Ακόμα δηλαδή και αν για ένα επιστήμονα έχουν εντοπιστεί πάνω από 1000 δημοσιεύσεις ή για μια δημοσίευση πάνω από 1000 citations, ο χρήστης έχει τη δυνατότητα να δει μόνο τα πρώτα 1000. Η κατάταξη καθορίζεται με βάση τον αριθμό των citations.

  ❖ Δεν επιτρέπει σε προγράμματα να αντλούν αυτόματα δεδομένα, για το λόγο αυτό άλλωστε δεν διατίθεται σχετικό API. Συνεπώς, όταν αντιληφθεί το Google Scholar ότι ο αποστολέας των queries είναι πρόγραμμα και όχι φυσικός χρήστης, αρνείται να απαντήσει στα επόμενα ερωτήματα (*HTTP status code : 403 Forbidden*)

Όσον αφορά τις κλάσεις του πακέτου αυτού, έχουν ως εξής:

- ***GSRejectionException***: το exception αυτό χρησιμοποιείται για να διακόψει τη μια on-line αναζήτηση στις περιπτώσεις που διαπιστώνεται ότι το Google Scholar αρνείται να απαντήσει στα σχετικά queries.
- ***GSParser***: η λειτουργία της κλάσης αυτής συνίσταται στο να απευθύνει ερωτήματα στο Google Scholar. Στη συνέχεια, ανακτά τις αντίστοιχες HTML σελίδες με τις απαντήσεις και τις διαχωρίζει σε υποσυμβολοσειρές, κάθε μια από τις οποίες αντιστοιχεί σε μια δημοσίευση. Χρησιμοποιώντας τις μεθόδους της *επόμενης κλάσης*, εξάγει πληροφορίες για όσα πεδία διαθέτουν οι εγγραφές αυτές και με τα δεδομένα αυτά δημιουργεί και ένα νέο αντικείμενο της Paper (προηγούμενο πακέτο). Να σημειώσουμε ότι η κλάση αυτή αναλαμβάνει να επεξεργαστεί δημοσιεύσεις οποιασδήποτε κατηγορίας, είτε το σχετικό αίτημα προκύπτει από την αναζήτηση των δημοσιεύσεων ενός συγγραφέα είτε από την αναζήτηση citations μιας δημοσίευσης
- ***FieldsParser***: η κλάση αυτή περιλαμβάνει το σύνολο των μεθόδων που χρησιμοποιούνται για την εξαγωγή πληροφορίας για κάθε πεδίο μιας δημοσίευσης, είτε από τις εγγραφές BibTex είτε από τις HTML σελίδες.

- ***WrapperUtilities***: στην κλάση αυτή έχουν τοποθετηθεί στατικές μέθοδοι γενικής χρησιμότητας για τη λειτουργία ενός wrapper

4.3.4 Μέθοδοι Μηχανικής Μάθησης (MachineLearning Package)

Το πακέτο αυτό περιλαμβάνει τις κλάσεις που εξασφαλίζουν την ορθή λειτουργία του συστήματος, καθώς υλοποιούν τις μεθόδους τεχνητής νοημοσύνης που παραθέσαμε στο προηγούμενο κεφάλαιο για την επίλυση των προβλημάτων citation matching, mixed και split citation. Για την ακρίβεια, οι κλάσεις αυτές έχουν ως εξής:

- ***ClassificationCMP***: σκοπός της κλάσης αυτής είναι η επίλυση του προβλήματος Citation Matching. Υλοποιεί δηλαδή τον αλγόριθμο που προσπαθεί να εντοπίσει ποιες από τις δημοσιεύσεις που προέκυψαν από μια on-line αναζήτηση πιθανόν αναφέρονται στην ίδια δημοσίευση ή ταυτίζονται με δημοσιεύσεις ήδη καταχωρημένες στη βάση δεδομένων.
- ***ClusteringMSCP***: η κλάση αυτή υλοποιεί τη μέθοδο ομαδοποίησης που επιλύει ταυτόχρονα τα προβλήματα Mixed Citation και Split Citation. Εντοπίζει δηλαδή το σύνολο των διαφορετικών επιστημόνων που συμμετέχουν τόσο στις δημοσιεύσεις της on-line αναζήτησης όσο και σε αυτές που είναι καταχωρημένες στη ΒΔ.
- ***AuthorVertex***: μέσω της κλάσης αυτής εμπλουτίζεται με τα απαραίτητα πεδία η κλάση UndirectedSparseVertex της βιβλιοθήκης JUNG ([OFSW]) της Java. Πιο συγκεκριμένα, τα αντικείμενά της αποτελούν τους κόμβους του γράφου των επιστημόνων που δημιουργείται και αναλύεται με την JUNG κατά την επίλυση των προβλημάτων mixed και split citation.
- ***SynonymSearch***: η κλάση αυτή υλοποιεί μια μέθοδο για τον εντοπισμό συνωνύμων του δοσμένου επιστήμονα, μετά από την ολοκλήρωση μιας on-line αναζήτησης. Επιχειρεί δηλαδή μια μερική λύση του split citation χρησιμοποιώντας τους τίτλους των δημοσιεύσεων που εντοπίζει μια on-line αναζήτηση.
- ***Constants***, ***Utilities***: ο ρόλος των κλάσεων αυτών είναι βοηθητικός για τη λειτουργία των μεθόδων του πακέτου και πανομοιότυπος με τις αντίστοιχες των προηγούμενων πακέτων.

5

Επίλογος

5.1 Σύνοψη και Συμπεράσματα

Στα πλαίσια της εργασίας αυτής είδαμε ότι η αυτόματη εύρεση των βιβλιογραφικών αναφορών θα διευκόλυνε σε σημαντικό βαθμό την αξιολόγηση του ερευνητικού έργου των επιστημόνων. Ωστόσο, οι μέχρι τώρα προσπάθειες, αν και αρκετά αξιόλογες και φιλόδοξες, δεν έχουν καταφέρει να λύσουν μια σειρά σχετικών προβλημάτων, με σημαντικότερα το citation matching, το mixed citation και το split citation. Στόχος μας ήταν η ανάπτυξη ενός συστήματος που παρέχει *αξιόπιστη* ανάλυση βιβλιογραφικών αναφορών, αντιμετωπίζοντας αυτά ακριβώς τα προβλήματα.

Η προσέγγισή μας συνοψίζεται στα ακόλουθα βήματα:

- Εξετάσαμε τις ακαδημαϊκές μηχανές αναζήτησης, διαπιστώνοντας ότι οι σημαντικότερες από αυτές (Scopus, Web of Science, Google Scholar) είναι συμπληρωματικές και όχι ανταγωνιστικές. Επιλέξαμε παρόλα αυτά να εργαστούμε μόνο με το Google Scholar, χωρίς να αποκλείουμε μελλοντική εκμετάλλευση και των υπολοίπων.
- Διερευνήσαμε τους τρόπους ανάπτυξης και συντήρησης ενός προγράμματος που εξάγει πληροφορίες από HTML σελίδες, από τη μορφή δηλαδή στην οποία παράγονται τα αποτελέσματα του Google Scholar. Αποφασίστηκε και οι δυο εργασίες να γίνονται χειρωνακτικά, επειδή τα εργαλεία που τις αυτοματοποιούν αποδείχτηκαν ανεπαρκή και ο φόρτος εργασίας είναι μικρός όσον αφορά μόνο ένα Web site, δηλαδή αυτό του Google Scholar.
- Μελετήσαμε τους αλγορίθμους που προτείνονται στη σχετική βιβλιογραφία για την επίλυση των προβλημάτων citation matching, mixed citation και split citation. Διαπιστώσαμε ομοφωνία όσον αφορά τις ιδιότητες στις οποίες βασίζουν οι αλγόριθμοι αυτοί τη λειτουργία τους. Στις ίδιες ιδιότητες στηρίχθηκε και η δική μας προσπάθεια. Ωστόσο, κανένας από τους προτεινόμενους αλγορίθμους δεν κρίθηκε κατάλληλος για

την εφαρμογή μας με αποτέλεσμα να αναπτύξουμε πρωτότυπους αλγορίθμους που ταυτόχρονα είναι καλύτερα προσαρμοσμένοι στις ανάγκες του συστήματός μας.

- Συγκρίναμε τις επιδόσεις των πιο διαδεδομένων μετρικών μορφολογικής απόστασης συμβολοσειρών για να διαλέξουμε ποια θα χρησιμοποιήσουμε στους αλγορίθμους μας. Σε γενικές γραμμές, καταλήξαμε στις Jaro και SoftTFIDF.
- Τέλος, προχωρήσαμε στην υλοποίηση του συστήματος με βάση το μοντέλο πρωτοτυποποίησης της τεχνολογίας λογισμικού. Έτσι, αρχικά σχεδιάσαμε και δημιουργήσαμε τη βάση δεδομένων και ένα πλήρως λειτουργικό πρωτότυπο, με τη βοήθεια του οποίου εξάγαμε τις λειτουργικές απαιτήσεις της εφαρμογής μας. Στη συνέχεια, σχεδιάσαμε και υλοποιήσαμε το τελικό σύστημα, με βάση την εμπειρία του πρωτότυπου. Η γλώσσα προγραμματισμού επιλέξαμε να είναι η Java, ενώ ως περιβάλλον διαχείρισης της βάσης δεδομένων χρησιμοποιήσαμε τη mySQL.

Έχοντας ολοκληρώσει την προσπάθεια αυτή, μπορούμε με ασφάλεια να πούμε ότι επιτεύχθηκαν όλοι οι στόχοι που αρχικά είχαν τεθεί. Πιο συγκεκριμένα, η εφαρμογή που αναπτύχθηκε διαθέτει ένα γραφικό περιβάλλον ιδιαίτερα εύχρηστο και φιλικό προς τον χρήστη, ώστε να μπορεί να τη χειριστεί άμεσα, χωρίς να απαιτείται σημαντική προηγούμενη εξοικείωση. Παρέχει επίσης δυνατότητες που καλύπτουν το σύνολο των λειτουργικών απαιτήσεων που καθορίσαμε παραπάνω, ενώ επίσης η δομή της καθιστά εύκολη τόσο την συντήρησή της όσο και την επέκτασή της με νέες δυνατότητες. Όσον αφορά στην επίδοση των αλγορίθμων μηχανικής μάθησης που αναπτύχθηκαν για τα προβλήματα citation matching, mixed citation και split citation, ελέγχθηκαν μόνο σε περιορισμένο εύρος δεδομένων. Σημαντικότερη αιτία για την αδυναμία αυτή είναι η πολιτική του Google Scholar, η οποία δεν επιτρέπει σε προγράμματα να αντλούν αυτόματα δεδομένα από αυτό. Συνεπώς, όταν αντιληφθεί ότι ο αποστολέας των ερωτημάτων είναι πρόγραμμα και όχι φυσικός χρήστης, αρνείται να απαντήσει στα ερωτήματα. Πάντως, από τις περιορισμένες δοκιμές που καταφέραμε να πραγματοποιήσουμε, διαπιστώσαμε ότι η απόδοσή των αλγορίθμων ήταν ικανοποιητική.

5.2 Μελλοντικές Επεκτάσεις

Παρόλο που τα πρώτα αποτελέσματα τις προσπάθειας μας είναι ικανοποιητικά, υπάρχει πλήθος θεμάτων που χρήζουν περαιτέρω διερεύνησης και βελτίωσης. Στη συνέχεια παραθέτουμε τα σημαντικότερα από τα θέματα, στα οποία θα επικεντρωθούμε στο μέλλον:

- Κατ' αρχήν σημαντικά θα ήταν τα οφέλη για τη χρηστικότητα της εφαρμογής από μια διεύρυνση των πηγών πληροφοριών της, προσθέτοντας δηλαδή τη δυνατότητα

αναζήτησης και από άλλες ακαδημαϊκές μηχανές αναζήτησης πέραν του Google Scholar. Εκείνές που σίγουρα αξίζει να συμπεριληφθούν σε αυτή τη διεύρυνση είναι το Scopus και το Web of Science. Συνδυάζοντας τα αποτελέσματα και των τριών αυτών υπηρεσιών, το σύστημά μας θα καλύπτει τη συντριπτική πλειοψηφία των on-line διαθέσιμων βιβλιογραφικών αναφορών, κάτι που σε όρους εξαγωγής πληροφορίας μεταφράζεται σε μεγιστοποίηση (σχεδόν) του recall.

- Περαιτέρω βελτιώσεις πρέπει να γίνουν και στους αλγορίθμους μηχανικής μάθησης που αναπτύξαμε. Ειδικότερα, εν όψει της ενσωμάτωσης αποτελεσμάτων από επιπρόσθετες μηχανές αναζήτησης, κρίνεται απαραίτητη η αναθεώρηση του αλγορίθμου ομαδοποίησης για το citation matching. Ο αλγόριθμος αυτός αναπτύχθηκε είναι προσαρμοσμένος στα αποτελέσματα του Google Scholar, ενώ, επιπλέον, χρησιμοποιεί για τις συγκρίσεις του σχετικά λίγα πεδία. Είναι αναγκαίο ωστόσο να αναπτύξουμε ένα γενικότερο ταξινομητή (classifier), ο οποίος θα συγκρίνει περισσότερα πεδία όσον αφορά τις βιβλιογραφικές αναφορές και τα βάρη του οποίου θα εκπαιδευθούν με βάση ένα προσεκτικά επιλεγμένο σύνολο δεδομένων. Ειδικότερα, όσον αφορά στον αλγόριθμο επίλυσης των προβλημάτων mixed και split citation, σημαντικές βελτιώσεις στα αποτελέσματά του θα προκύψουν, με βάση πάντα τη βιβλιογραφία, αν τον μετατρέψουμε έτσι ώστε :
  - να εξετάζει τη σημασιολογική ταύτιση των τίτλων των δημοσιεύσεων, αντί για την απλή ταύτιση λέξεων που εξετάζει στη παρούσα του μορφή,
  - να αποδίδει κατά τη σύγκριση των URL δυο δημοσιεύσεων βάρη αντιστρόφως ανάλογα της συχνότητας τους. Δηλαδή η ταύτιση δυο "σπάνιων" domains, www.ntua.gr για παράδειγμα, να έχει μεγαλύτερη αξία από την ταύτιση "συχνών" domains, όπως του portal.acm.org.

  Επίσης, τα αποτελέσματά των δυο αυτών αλγορίθμων μηχανικής μάθησης θα βελτιώνονταν αν προσθέταμε τη δυνατότητα για αυτόματη αναζήτηση πρόσθετων δεδομένων σχετικά με δυο δημοσιεύσεις ή δυο συγγραφείς για τους οποίους ο αντίστοιχος αλγόριθμος δεν καταλήγει σε απόλυτα ασφαλή συμπεράσματα.
- Τέλος, βελτιώσεις επιδέχεται το σύστημα και όσον αφορά την αλληλεπίδραση με το χρήστη, δηλαδή το γραφικό περιβάλλον και τις δυνατότητες που αυτό παρέχει στο χρήστη. Πιο συγκεκριμένα, κρίνεται άκρως απαραίτητη η προσθήκη της δυνατότητας για αναζήτηση των νέων citations για μια συγκεκριμένη δημοσίευση, αναζήτηση η οποία στην παρούσα φάση γίνεται μόνο έμμεσα (μέσω της αναζήτησης των δημοσιεύσεων ενός από τους συγγραφείς της). Πρέπει, επίσης, να γίνει περισσότερο ευέλικτη η συγχώνευση των δημοσιεύσεων αλλά και να εμπλουτιστεί η εφαρμογή με τη δυνατότητα μεταβολής από το χρήστη των δεδομένων που αφορούν μια δημοσίευση, τόσο πριν όσο και μετά από την καταχώρησή της στη βάση δεδομένων.

Επιπλέον, σημαντική θα ήταν η προσθήκη για δυνατότητα οπτικοποίησης του κοινωνικού δικτύου που δημιουργείται μέσω των citations μεταξύ των επιστημόνων. Τέλος, θα ήταν ενδιαφέρον να μετατραπεί η εφαρμογή αυτή σε διαδικτυακή υπηρεσία.

Βιβλιογραφία

Ελληνική Βιβλιογραφία

[Ραπ07] Σ. Ραπανάκης, "*Αξιολόγηση μετρικών μορφολογικής απόστασης λέξεων και εφαρμογή τους στην ταύτιση ομοίων εγγραφών*", Διπλωματική Εργασία, Ιούνιος 2007

[Συγ05] Γ.Β. Συγλέτος, "*Εξόρυξη γνώσης για εξαγωγή πληροφορίας από τον παγκόσμιο ιστό με χρήση τεχνικών ψηφοφορίας και συσσωρευμένης γενίκευσης*', Διδακτορική Διατριβή, Νοέμβριος 2005

[Τσου06] Χ.Ε. Τσουρακάκης, "*Μέθοδοι αυτόματου εντοπισμού σφαλμάτων και βελτίωσης wrappers με χρήση επαυξητικών μεθόδων μάθησης*", Διπλωματική Εργασία, Νοέμβριος 2006

Ξένη Βιβλιογραφία

[BMR06] J. Bosman, I. Mourik, M. Rasch, E. Sieverts, H.Verhoeff, "*Scopus reviewed and compared*", Utrecht University Library, 2006

[Broo86] T.A. Brooks, "*Evidence of complex citer motivations*", Journal of the American Society for Information Science, v. 37, pp.34-36, 1986

[Burr06] M. Burright, "*Google Scholar – Science & Technology*", 2006, available at: http://www.istl.org/06-winter/databases2.html

[CRF03] W. Cohen, P. Ravikumar, S.Fienberg, "*A Comparison of String Distance Metrics for Name-Matching Tasks*", In IIWeb Workshop held in conjunction with IJCAI, 2003

[Dess06] H.Dess, "*Databases Reviews & Reports - Scopus*", 2006, available at: http://www.istl.org/06-winter/databases4.html

[Fing06] S. Fingerman, "*Electronic Resources Reviews – Web of Science and Scopus: Current Features and Capabilities*", 2006, available at: http://www.istl.org/06-fall/electronic2.html

[GBL98] C. Giles, K. Bollacker, S. Lawrence, "*Citeseer: An Automatic Citation Indexing system*", In Digital Libraries 98 – The Third ACM Conference on Digital Libraries, pp. 89-98, 1998

[Godi07] Lluís Codina, "*Search engines for scientific and academic information*" [on line]. "Hipertext.net", num. 5, 2007, available at : http://www.hipertext.net/english/pag1021.htm.

[GR02] J. Grabmeier, A. Rudolph, "*Techniques of Cluster Algorithms in Data Mining*", Data Mining and Knowledge Discovery, 6, pp. 303-360, 2002.

[HGZ04] H. Han, L. Giles, H. Zha, C. Li, K. Tsioutsiouliklis, "*Two Supervised Learning Approaches for Name Disambiguation in Author Citations*", JCDL, 2004

[HXZG05] H. Han, W. Xu, H. Zha, C. Giles, "*A Hierarchical Naive Bayes Mixture Model for Name Disambiguation in Author Citations*", SAC, 2005

[HZG05] H. Han, H. Zha, C. Giles, "*Name Disambiguation in Author Citations using a K-way Spectral Clustering Method*", JCDL, 2005.

[Jasc05] P. Jasco, "*Google Scholar and The Scientist*", available at: http://www2.hawaii.edu/~jacso/extra/gs/

[Jasc06] P. Jasco, "*Google Scholar, the mis-matchmaker*", available at: http://www2.hawaii.edu/~jacso/extra/GS-mis-matchmaker/

[KMP06] P. Kanani, A. McCallum, C. Pal, "*Improving Author Coreference by Resource-bounded Information Gathering from the Web*", Technical note, 2006

[LOPK05] D. Lee, B.W. On, J. Kang, S. Park, "*Effective and Scalable Solutions for Mixed and Split Citation Problems in Digital Libraries*", In IQIS 2005.

[LGB99] S. Lawrence, C. Lee Giles, Kurt Bollacker, "*Digital Libraries and Autonomous Citation Indexing*", IEEE Computer, v. 32, no. 6, pp. 67-71, 1999

[LRST02] A. Laender, B. Ribeiro-Neto, A. da Silva, J. Teixeira. "*A Brief Survey of Web Data Extraction Tools*", SIGMOD Record, 31(2), 2002.

[MA05] J. Marinacci, C. Adamson, "*Swing Hacks*", O'Reilly Media ©, 2005

[Mit97] T. Mitchell, "*Machine Learning*", McGraw-Hill, 1997

[Mit06] T. Mitchell, "*Generative and Discriminative classifiers: Naive Bayes and Logistic Regression*", draft available at: www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf

[MSS06] D. McRae-Spencer, N. Shadbolt, "*Also By The Same Author: AKTiveAuthor, a Citation Graph Approach to Name Disambiguation*", JCDL, 2006

[OFSW] J. O'Madadhain, D. Fisher, P. Smyth, S. White, Y. Boey, "*Analysis and Visualization of Network Data using JUNG*", Journal of Statistical Software, volume 5, issue 2

[OLK05] B. On, D. Lee, J. Kang, P. Mitra, "*Comparative Study of Name Disambiguation Problem using a Scalable Blocking-based Framework*", JCDL, 2005

[PMMR03] H. Pasula, B. Marthi, B. Milch, S. Russel, I. Shpitser, "*Identity uncertainty and citation matching*", Advances in Neural Information Processing (NIPS), 2003

[Shim] M. Shimbo, "*A Robust Method for Citation Matching*"

[SKS02] A. Silberschatz, H. Korth, S. Sudarshan, "*Database Systems Concepts*", The McGraw-Hill Company Inc, 2002

[SP06] K. Stergiou, D. Pauly, "*Equivalence of results from two citation analyses: Thomson ISI's Citation Index and Google's Scholar service*", Ethics in Science and Environmental Politics, 2005

[TKL06] Y. Tan, M. Kan, D. Lee, "*Search Engine Driven Author Disambiguation*", JCDL, 2006

[WI07] Wikipedia, "*Wikipedia, the free encyclopedia*", available at: http://en.wikipedia.org

[WMPH04] B. Wellner, A. McCallum, F. Peng, M. Hay, "*An Integrated, Conditional Model of Information Extraction and Coreference with Application to Citation Matching*", Proceedings of the 20th conference on Uncertainty in artificial intelligence, pp. 593-601, 2004

[Zill07] M. Zillman, "*Academic and Scholar Search Engines and Sources-An Internet MiniGuide Annotated Link Compilation*", available at: http://www.whitepapers.us/