Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Μεταπτυχιακή Διατριβή
στα Πληροφοριακά και Επικοινωνιακά Συστήματα

Δημιουργία Συνόλων από Δεδομένα για Χρήση Αξιολόγησης Αλγορίθμων Ταύτισης Αντικειμένων

Μάριος Νικολάου

Επιβλέπων Καθηγητής
Αικατερίνη Ιωάννου

Μάιος 2017

Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Δημιουργία Συνόλων από Δεδομένα για Χρήση Αξιολόγησης Αλγορίθμων Ταύτισης Αντικειμένων

Μάριος Νικολάου

Επιβλέπων Καθηγητής
Αικατερίνη Ιωάννου

Η παρούσα μεταπτυχιακή διατριβή υποβλήθηκε
προς μερική εκπλήρωση των απαιτήσεων για απόκτηση

μεταπτυχιακού τίτλου σπουδών
στα Πληροφοριακά Συστήματα

από τη Σχολή Θετικών και Εφαρμοσμένων Επιστημών
του Ανοικτού Πανεπιστημίου Κύπρου

Μάιος 2017

Περίληψη

Το πρόβλημα της ταυτοποίησης αντικειμένων συνεχίζει να προβληματίζει την ερευνητική κοινότητα. Με την ραγδαία εξάπλωση του διαδικτύου δεν μπορεί να ελεγχτεί η καταχώρηση πληροφοριών σε βάσεις δεδομένων με αποτέλεσμα την δημιουργία εγγραφών που αναφέρονται είτε στο ίδιο πρόσωπο, είτε στην ίδια ταινία, είτε στο ίδιο άρθρο κτλ. αλλά με διαφορετικό τρόπο. Αυτός είναι και ο ορισμός της ταυτοποίησης αντικειμένων. Πολλοί ερευνητές έχουν δημιουργήσει αλγόριθμους με σκοπό την ταύτιση αντικειμένων. Το πρόβλημα όμως είναι η έλλειψη συνόλων από δεδομένα για την αξιολόγηση αυτών ώστε να μπορούν να χρησιμοποιηθούν ευρέως. Στόχος της μεταπτυχιακής διατριβής είναι η δημιουργία τέτοιων συνόλων δεδομένων για χρήση αξιολόγησης αλγορίθμων ταύτισης αντικειμένων. Αρχικά έγινε ο καθορισμός των οικογενειών των συνόλων δεδομένων και στην συνέχεια δημιουργήθηκαν. Χρησιμοποιήθηκαν δεδομένα και ιδιότητες που συνήθως χρησιμοποιούνται στην ερευνητική κοινότητα. Στην συνέχεια χρησιμοποιώντας το σύστημα EMBench έγινε η δημιουργία των οικογενειών. Η κάθε οικογένεια περιέχει συλλογές όπου κάθε μια διαφέρει σε κάποιες παραμέτρους ή στον αριθμό των παρατηρήσεων, έτσι ώστε να υπάρχουν σύνολα δεδομένων που μπορούν καλύπτουν ευρέως την αξιολόγηση ενός αλγορίθμου ταύτισης αντικειμένων.

Summary

The problem of entity matching continues to concern the research community. With the rapid spread of the Internet, it is not possible to control the entry of information into databases, resulting in the creation of records referring to either the same person, either in the same movie or in the same article, but in a different way. This is the definition of entity matching. Many researchers have created algorithms for entity matching. The problem, however, is the lack of data sets to evaluate them so they can be widely used. The aim of this postgraduate dissertation is the creation of such data sets for the use of evaluation of entity matching algorithms. Initially, the families of the data sets will be defined and then will be created. The data sets and the attributes were used, are commonly used in the research community. Then, through experiments using the EMBench system, families were created. Families were divided into collections each differing in some parameters or the number of observations so that there will be datasets that can broadly cover the evaluation of an entity matching algorithm.

Ευχαριστίες

Ένα μεγάλο ευχαριστώ στην επιβλέπων καθηγήτρια μου , κυρία Αικατερίνη Ιωάννου, Λέκτορα του ΑΠΚυ. Καταρχάς, την ευχαριστώ για την ευκαιρία που μου έδωσε να συνεργαστώ μαζί της για την εκπόνηση της μεταπτυχιακής διατριβής και κατά επέκταση για όλη την υποστήριξη και καθοδήγηση σε αυτή τη μεγάλη διαδρομή. Επίσης θέλω να ευχαριστήσω την σύζυγο μου Νατάσα και την κόρη μου Έμιλυ για την υπομονή που έδειξαν για όλες αυτές τις ώρες που ήμουν αφοσιωμένος στην εκπόνηση της μεταπτυχιακής διατριβής και για όλα τα Σαββατοκύριακα που έπρεπε να απουσιάζουν από το σπίτι για να μπορώ να συγκεντρωθώ στην εκπόνηση της μεταπτυχιακής διατριβής. Τέλος, ευχαριστώ τα δύο μεγαλύτερα μου αδέλφια , Στέλλα και Άντρο, που με την ηθική τους στήριξη , την εμπειρία αλλά και τις συμβουλές τους, μου έδιναν δύναμη ώστε να καταφέρω να διεκπεραιώσω την μεταπτυχιακή διατριβή.

Περιεχόμενα

Κεφάλαιο 1
Εισαγωγή

Κατά την διάρκεια των τελευταίων ετών, πολλές εφαρμογές ανταλλάσσουν δεδομένα ή χρησιμοποιούν δεδομένα που υπάρχουν διαθέσιμα στον Παγκόσμιο Ιστό. Ωστόσο, λόγω διαφόρων υποκειμενικών και αντικειμενικών λόγων όπως ελλιπών πληροφοριών, λανθασμένης ορθογραφίας και καταχώρησης μη ομοιόμορφων στοιχείων στις εγγραφές, οι αποθηκευμένες πληροφορίες σε πηγές δεδομένων είναι μη ακριβείς και ατελείς. Αποτέλεσμα αυτών των συνθηκών είναι να δημιουργείται ένα μεγάλο ζήτημα για τη ποιότητα των πληροφοριών αφού μια πληροφορία που αναφέρεται στο ίδιο αντικείμενο θα υπάρχει σε πηγές πληροφοριών διαφορετικά καταχωρημένη. Η κακή ποιότητα πληροφοριών μπορεί να βλάψει την αποτελεσματικότητα των εργασιών ανάλυσης και εξέτασης σε πολλές εφαρμογές, έτσι είναι απαραίτητο να βρεθούν οι κατάλληλες λύσεις για τον καθαρισμό και την επιδιόρθωση λανθασμένων πληροφοριών[CJZ+12].

Σημαντικό πρόβλημα σε κάθε ενσωμάτωση πληροφορίας και εφαρμογής καθαρισμού πληροφοριών είναι η ικανότητα να αναγνωρίζεται αν δυο διαφορετικές παρατηρήσεις (ο όρος παρατήρηση υποδεικνύει κάθε καταχώρηση ξεχωριστά, που περιλαμβάνει εγγραφές σε ένα σύνολο δεδομένων) πληροφοριών παρουσιάζουν το ίδιο πραγματικό αντικείμενο, π.χ. μια ταινία, ένα βιβλίο, ένα πρόσωπο. Αυτή η εργασία είναι τυπικά γνωστή ως ταυτοποίηση αντικειμένων, αλλά στην βιβλιογραφία την βρίσκουμε και ως deduplication, entity resolution, record linkage, object identification ή merge-purge [BGM+09], [IRV13], [LLH14]. Αυτό το οποίο κάνει δύσκολη την διαδικασία ταυτοποίησης είναι η ετερογένεια που υπάρχει στα σύνολα δεδομένων επειδή είτε έχουν δημιουργηθεί από διαφορετικές εφαρμογές, είτε επειδή υπάρχουν εκ φύσεως σφάλματα και ασυνέπειες σε αυτά, είτε επειδή κατά τον σχεδιασμό και την ανάπτυξη τους, οι ειδικοί είχαν υπόψη διαφορετικές απαιτήσεις για κάθε σύνολο δεδομένων. Για τον λόγο αυτό πιθανό να υπάρχουν ατέλειες και λάθη στις πληροφορίες[IRV13].

Η εν λόγω τακτική οδήγησε στην δημιουργία διαφορετικών περιγραφών για τα ίδια αντικείμενα. Η ερευνητική κοινότητα προσπάθησε να επιλύσει το συγκεκριμένο πρόβλημα με τον σχεδιασμό και ανάπτυξη καινοτόμων αλγορίθμων για ταύτιση αντικειμένων καθώς και συστημάτων για την παραγωγή συνόλων δεδομένων και την αξιολόγηση αλγορίθμων. Ένα από θέματα που καλούνται να αντιμετωπίσουν οι ερευνητές που ασχολούνται με αυτού του είδους τους αλγορίθμους είναι η έλλειψη δεδομένων για τη διεξαγωγή πειραμάτων αξιολόγησης των αλγορίθμων που έχουν υλοποιήσει για την ταύτιση αντικειμένων.

Η ταυτοποίηση αντικείμενων (entity matching) είναι μια ενδιαφέρον διαδικασία ειδικά για αντικείμενα τα οποία είναι ετερογενή και με περιορισμένη ποιότητας πληροφοριών [IRV13], [KR10]. Το πρόβλημα ταυτοποίησης αρχικά ορίστηκε από τον Newcombe και άλλους το 1959 και τυποποιήθηκε από τους Felligi και Sunter δέκα χρόνια αργότερα[KTR09], [KR10]. Έχουν προταθεί πολλές προσεγγίσεις για την ταυτοποίηση αντικειμένων ειδικά για ελεγχόμενες και σχεδιασμένες πληροφορίες (βλέπε τμήμα 2.1).

Η ταυτοποίηση αντικειμένων θεωρείται σημαντική για την ερευνητική κοινότητα. Αν βρεθεί τρόπος ή μέθοδος ώστε να γίνεται ταυτοποίηση αντικειμένων σε τέτοιο ποσοστό που να θεωρείται βέλτιστο, τότε θα εξαλειφτούν και οι ελλιπείς πληροφορίες που υπάρχουν στις πηγές συνόλων δεδομένων. Για να γίνει αυτό θα πρέπει παράλληλα να μετριαστεί η λανθασμένη καταχώρηση πληροφοριών, ώστε να αποφευχθεί η αύξηση της καταχώρησης λανθασμένων και ελλιπών πληροφοριών.

Αυτή η μεταπτυχιακή διατριβή περιέχει δύο βασικές συνεισφορές. Η πρώτη είναι ο καθορισμός των οικογενειών με βάση τα πειράματα που συνήθως εκτελούνται στις υπάρχουσες δημοσιεύσεις. Η δεύτερη είναι η δημιουργία των οικογενειών καθώς και των συνόλων της κάθε οικογένειας.

1.1 Αναγκαιότητα και Σπουδαιότητα Έρευνας

Τα υφιστάμενα σύνολα για ταύτιση αντικειμένων καθώς και τα συστήματα για παραγωγή συνόλων έχουν αρκετά προβλήματα που επιφέρουν περιορισμούς στους ερευνητές. Αυτά τα προβλήματα περιλαμβάνουν τον μη επαρκή μεγάλο αριθμό δείγματος, το μικρό αριθμό ιδιοτήτων και μη σαφή ετερογένεια καθώς και μια σημαντική παράμετρο τη μη συντακτική παραλλαγή. Η δημιουργία καινούργιων οικογενειών συνόλων δεδομένων, θα βοηθήσει την ερευνητική κοινότητα στην

αξιολόγηση αλγορίθμων ταύτισης αντικειμένων, ώστε να λυθεί το πρόβλημα των διαφορετικών περιγραφών των ιδίων αντικειμένων. Ο καθορισμός ελεγχόμενων οικογενειών συνόλων δεδομένων, είναι απαραίτητος, ώστε να δημιουργηθούν οικογένειες συνόλων δεδομένων που συνήθως χρησιμοποιούνται για την διεξαγωγή πειραμάτων για αξιολόγηση αλγορίθμων ταύτισης αντικειμένων.

Τα υπάρχουσα σύνολα δεδομένων ίσως να έχουν δημιουργηθεί μέσω της συνάθροισης πληροφοριών περισσότερων από ένα είδος ετερογένειας σε βαθμό που δεν είναι σαφές ακόμη και σε έναν έμπειρο χρήστη τι είναι αυτές οι ετερογένειες. Είναι σημαντικό να δημιουργηθεί και να παρέχεται ένα σύνολο δεδομένων το οποίο χαρακτηρίζεται από μια ελεγχόμενη σαφή ποικιλία διαφορετικών ετερογενειών μεταξύ των δεδομένων.

Το σύστημα που χρησιμοποιήθηκε για τη παραγωγή των οικογενειών συνόλων δεδομένων είναι το EMBench [IRV13], [IV14]. Το EMBench είναι ένα σύστημα αρχών για την αξιολόγηση των συστημάτων ταύτισης αντικειμένων. Προσφέρει μια μοναδική υπόθεση δοκιμής προσέγγισης η οποία συνδυάζει διαφορετικά επίπεδα τύπων, πολυπλοκότητας και κλιμάκων, επιτρέποντας μια πλήρη και ακριβή αξιολόγηση για τις διάφορες πτυχές ενός συστήματος ταυτοποίησης.

Σκοπός της μεταπτυχιακής διατριβής είναι ο καθορισμός των οικογενειών βάση διαφόρων πειραμάτων και ετερογενειών τα οποία θα συλλεχθούν μέσα από υπάρχουσες δημοσιεύσεις που υπάρχουν στην ερευνητική κοινότητα. Επιπλέον θα γίνει δημιουργία συνόλων από δεδομένα διαφορετικής ποικιλίας και μεγαλύτερου αριθμού δείγματος, μέσα από το σύστημα EMBench, τα οποία θα μπορούν να χρησιμοποιηθούν για την αξιολόγηση αλγορίθμων. Μετέπειτα τα σύνολα θα ενταχθούν σε οικογένειες έτσι ώστε η κάθε οικογένεια να προσφέρει αξιολόγηση για κάποιο συγκεκριμένο θέμα. Ως αποτέλεσμα αυτές οι οικογένειες θα αποτελούν βασικό βοήθημα στους ερευνητές για να μπορούν να αξιολογούν την ποιότητα και την συμπεριφορά κάποιου αλγόριθμου όταν ο αριθμός των αντικειμένων μετατρέπεται ή κάποια παράμετρος διαφοροποιείται.

1.2 Βασικά Ερευνητικά Ερωτήματα

Με βάση τα πιο πάνω, η εν λόγω μεταπτυχιακή διατριβή έχει σκοπό να μελετήσει ερευνητικά ερωτήματα που σχετίζονται με την αξιόλογης αλγορίθμων ταύτισης αντικειμένων. Αυτά τα ερευνητικά ερωτήματα δίνονται στην ακόλουθη λίστα και αναλύονται στα κεφάλαια που ακολουθούν:

- Τι δεδομένα χρησιμοποιούν οι ερευνητές για να αξιολογήσουν τους αλγορίθμους ταύτισης που έχουν υλοποιήσει;
- Ποιες οικογένειες και σύνολα είναι χρήσιμο να δημιουργηθούν για τους σκοπούς της αξιολόγησης;
- Πως κάποιος ερευνητή μπορεί να χρησιμοποιήσει τα σύνολα που δημιουργήθηκαν στην εν λόγω μεταπτυχιακή διατριβή;

1.3 Σύνοψη Μεταπτυχιακής Διατριβής

Στην παρούσα μεταπτυχιακή διατριβή μέσα από σχετική βιβλιογραφία θα παρουσιαστεί το πρόβλημα ταύτισης αντικειμένων και πως διάφοροι ερευνητές/συγγραφείς προσπάθησαν να το εξαλείψουν. Γίνετε αναφορά σε διάφορα συστήματα που έχουν χρησιμοποιηθεί για την δημιουργία συνόλων δεδομένων για χρήση αξιολόγησης αλγορίθμων ταύτισης αντικειμένων καθώς επίσης αναφέρονται οι διάφορες τεχνικές ταυτοποίησης (Κεφάλαιο 2). Περεταίρω στο Κεφάλαιο 3.1 γίνεται αναφορά στα υφιστάμενα σύνολα δεδομένων που έχουν χρησιμοποιηθεί σε υφιστάμενες μελέτες και στο Κεφάλαιο 3.2 οι υπάρχουσες μελέτες που έχουν δημιουργήσει σύνολα και τον τρόπο που έχουν δημιουργηθεί. Στην συνέχεια γίνετε καθορισμός των οικογενειών που θα δημιουργηθούν στο Κεφάλαιο 3.3. Ο καθορισμός αυτών βασίστηκε σε υπάρχουσες βιβλιογραφικές αναφορές όπου έγινε χρήση των συγκεκριμένων δεδομένων. Για την δημιουργία αυτών των οικογενειών συνόλων δεδομένων θα χρησιμοποιηθεί το σύστημα EMBench που αναφέρεται στο Κεφάλαιο 4. Κάνοντας τις κατάλληλες ρυθμίσεις και χρησιμοποιώντας διάφορες παραμέτρους στο τεμαχιστή, θα δημιουργηθούν οι οικογένειες συνόλων δεδομένων. Αυτές θα έχουν κάποιες συγκεκριμένες διαφοροποίησης μεταξύ τους, όπως τον αριθμό αντικειμένων, τον αριθμό ιδιοτήτων κ.α. Σκοπός είναι οι οικογένειες που θα δημιουργηθούν, να είναι με τέτοιο τρόπο δημιουργημένες ώστε να μπορούν χρησιμοποιηθούν για την αξιολόγηση αλγορίθμων ταύτισης αντικειμένων.

1.4 Δομή Μεταπτυχιακής Διατριβής

Το εναπομείναντα κείμενο είναι χωρισμένο σε έξι κεφάλαια. Στο δεύτερο κεφάλαιο, βιβλιογραφική επισκόπηση, αναλύονται σημαντικά θέματα όπως τι είναι ταυτοποίηση αντικειμένων, τεχνικές αντικειμένων και συστήματα ταυτοποίησης. Στο τρίτο κεφάλαιο παρουσιάζονται υφιστάμενες μελέτες που αφορούν σύνολα δεδομένων που συνήθως χρησιμοποιούνται καθώς και ποιες ιδιότητες. Επίσης στο κεφάλαιο θα γίνει καθορισμός των οικογενειών και των ιδιοτήτων αυτών, που θα δημιουργηθούν, που θα είναι στηρίζονται στις μέχρι τώρα μελέτες που έχουν γίνει. Στο κεφάλαιο τέσσερα γίνεται ανάλυση της μεθοδολογίας της δημιουργίας των συνόλων δεδομένων. Παρουσιάζεται το σύστημα EMBench το οποίο θα χρησιμοποιηθεί για την δημιουργία των συνόλων δεδομένων και τέλος γίνεται η δημιουργία των οικογενειών των συνόλων δεδομένων όπως αυτές έχουν καθοριστεί στο κεφάλαιο τρία. Στο κεφάλαιο πέντε γίνεται ανάλυση των συνόλων δεδομένων που δημιουργήθηκαν, δίνονται τα αποτελέσματα και γίνεται συζήτηση για κάθε οικογένεια που δημιουργήθηκε. Στο έξι γίνεται σύνοψη της όλης έρευνας και δίνεται ο επίλογος.

1.5 Ορολογίες

Entity matching: Διαδικασία ταυτοποίησης δεδομένων που αντιπροσωπεύουν το ίδιο αντικείμενο σε μια μόνο βάση δεδομένων ή διαφορετικών πηγών

EMBench: Σύστημα αρχών αξιολόγησης συστημάτων ταυτοποίησης αντικειμένων

Shredders: Τεμαχιστές

Modifier/Destructor:Τροποποιητής

Κεφάλαιο 2

Βιβλιογραφική Επισκόπηση

Το δεύτερο κεφάλαιο, η βιβλιογραφική επισκόπηση αναφέρεται στις κύριες έννοιες και θέματα τα οποία αφορούν την ταυτοποίηση αντικειμένων, τα συστήματα ταυτοποίησης και τις τεχνικές ταυτοποίησης αντικειμένων. Μέσα από τη βιβλιογραφική επισκόπηση δίνεται η ευκαιρία να επισημανθούν γενικές απόψεις από διάφορους συγγραφείς όσο αφορά το θέμα της έρευνας. Περαιτέρω, δίνεται η ευκαιρία να παρουσιαστεί λεπτομερής βιβλιογραφική έρευνα η οποία έχει διεξαχθεί για να παρουσιάσει τα θέματα τα οποία συλλέχθηκαν.

Για τη συλλογή βιβλιογραφίας χρησιμοποιήθηκαν 24 επιστημονικά άρθρα, χρονολογημένα από το 2000. Το κύριο θέμα των επιστημονικών άρθρων είναι η ταυτοποίηση αντικειμένων , τα συστήματα ταυτοποίησης, η αξιολόγηση αλγορίθμων και τα σύνολα δεδομένων που συνήθως χρησιμοποιούνται και οι ιδιότητες αυτών. Οι χώροι αναζήτησης και εύρεσης των απαραίτητων πληροφοριών ήταν κυρίως ηλεκτρονικές βιβλιοθήκες, Έγινε εύρεση βιβλίων ή άρθρων από συλλογές κάθε βιβλιοθήκης (αυτοματοποιημένος κατάλογος βιβλιοθήκης, My Athens) αλλά και άλλων βιβλιοθηκών που παρέχονται από το Ανοικτό Πανεπιστήμιο Κύπρου. Η αναζήτηση έγινε χρησιμοποιώντας λέξεις κλειδιά στα αγγλικά όπως entity matching, deduplication, EMBench,algorithm,data set,evaluation.

2.1 Ταυτοποίηση Αντικειμένων

Όπως ήδη έχει αναφερθεί στο πρώτο κεφάλαιο η ταυτοποίηση δεδομένων (entity matching) είναι μια ενδιαφέρον διαδικασία ειδικά για αντικείμενα τα οποία είναι ετερογενή και στερούν στη ποιότητα πληροφοριών [IRV13], [KTR09]. Για την ταυτοποίηση αντικειμένων υπάρχουν αλγόριθμοι οι οποίοι όταν τους δίνεται μια συλλογή δεδομένων και ένα αντικείμενο, ταυτοποιούν τα δεδομένα από τη συλλογή τα

οποία ταιριάζουν καλύτερο στο αντικείμενο ή παράγουν μια ταξινομημένη λίστα δεδομένων βασισμένη στο αποτέλεσμα ταυτοποίησης με το αντικείμενο [IRV13].

Η ταυτοποίηση αντικειμένων είναι η διαδικασία ταυτοποίησης δεδομένων που αντιπροσωπεύουν το ίδιο αντικείμενο σε μια μόνο βάση δεδομένων ή διαφορετικών πηγών [PIN+12],[LLH14],[KR10],[IRV13],[BG07]. Γενικά η ταυτοποίηση αντικειμένων είναι αυτό το οποίο εννοείται από τον όρο ταυτοποίηση υπόθεσης. Τα αντικείμενα έχουν γίνει ουσιώδης δομές στις εφαρμογές των σημασιολογικών εννοιών του δικτύου τα οποία έχουν αυξήσει το ενδιαφέρον της ερευνητικής και βιομηχανικής κοινότητας στην ταυτοποίηση αντικειμένων για την ετερογενή ενσωμάτωση πληροφοριών του δικτύου για να απαντηθούν ερωτήματα. Επιπλέον, η ταυτοποίηση αντικειμένων θεωρείται σημαντική στην εξέλιξη των αντικειμένων όπου οι ταυτοποιήσεις ομοιοτήτων, μέσα από διαφορετικές δομές πληροφοριών, ίσως επιδεικνύουν ότι οι δυο δομές αντιπροσωπεύουν διαφορετικές φάσεις διάρκειας ζωής ενός μόνο αντικειμένου [IRV13].

Η παραπάνω διαδικασία έχει τη δυνατότητα να χρησιμοποιηθεί για διαφορετικούς σκοπούς. Ένας σκοπός είναι η εξάλειψη περιττών πληροφοριών (deduplication), για παράδειγμα ο εντοπισμός σε μια συλλογή δεδομένων διαφόρων απεικονίσεων του ιδίου αντικειμένου και η συγχώνευση αυτών των απεικονίσεων έτσι ώστε να γίνει μια. Άλλος σκοπός είναι να βρίσκει δομές σε δυο διαφορικές πηγές οι οποίες επιδεικνύουν το ίδιο πραγματικό αντικείμενο και χρησιμοποιεί αυτή τη πληροφορία για συγχώνευση των πηγών ή την ανταλλαγή πληροφοριών. Η ταυτοποίηση αντικειμένων η οποία στοχεύει στο να βρίσκει παρόμοιες απεικονίσεις οι οποίες αναφέρονται στο ίδιο αντικείμενο, είναι κρίσιμη σε διάφορα πεδία, όπως καθαρισμό και ενσωμάτωση πληροφοριών [CJZ+12], [KTR09], [LLH14]. Τελευταίος σκοπός είναι η απάντηση ερωτήσεων. Μια απορία ενός χρήστης είναι μια λίστα προδιαγραφών η οποία περιγράφει τα επιθυμητά χαρακτηριστικά του αντικειμένου που ψάχνει ο χρήστης στην μορφή ζευγών των ιδιοτήτων με χαρακτηριστικά το όνομα και την αξία.[IRV13].

Κυρίως υπάρχουν δυο είδη μεθόδων ταυτοποίησης αντικειμένων, συμπεριλαμβάνοντας την μέθοδο ταξινόμησης και την μέθοδο κανόνα[CJZ+12]. Η μέθοδος κανόνα είναι η πιο δημοφιλής γιατί είναι προσβάσιμη και εξηγήσιμη. Τα δεδομένα είναι σημαντικά σε πολλούς τομείς, συμπεριλαμβανομένου της οικονομίας, βιομηχανίας, ιατρικής, τεχνολογίας, κλπ. Ωστόσο, λόγω υποκειμενικών και αντικειμενικών λόγων όπως ορθογραφικά λάθη ή παραλειπόμενες αξίες οι πληροφορίες που αποθηκεύονται στις

πηγές είναι ανακριβής ή ελλιπής. Επομένως είναι απαραίτητο να ληφθούν μέτρα για να γίνει καθαρισμός και επισκευή των λανθασμένων πληροφοριών. Μια σημαντική εργασία στον καθαρισμό πληροφοριών και ενσωμάτωσης είναι η ταυτοποίηση αντικειμένων η οποία στοχεύει στο να καταχωρήσει ζεύγη του ίδιου αντικειμένου στον πραγματικό κόσμο[CJZ+12], [KTR09].

Ερευνητικές προσπάθειες διεξάγονται διαρκώς για να διορθώσουν το πρόβλημα ταυτοποίησης αντικειμένων αλλά ακόμη φαίνεται πως η ταυτοποίηση είναι ένα μεγάλο ζήτημα για την ερευνητική κοινότητα. Αρχικά, δεν υπάρχει καμία εφαρμογή ταυτοποίησης η οποία να λειτουργεί αποτελεσματικά για τις εργασίες ταυτοποίησης αντικειμένων. Για κάθε τομέα ξεχωριστά πρέπει να υπάρξει η βέλτιστη λειτουργία εφαρμογής ταυτοποίησης από το σύνολο των εφαρμογών για να λειτουργήσει αποτελεσματικά[LLH14]. Παρ' όλες τις διάφορες τεχνικές για την ταυτοποίηση αντικειμένων δεν υπάρχει καμία μεθοδολογία αξιολόγησης η οποία να καλύπτει όλες τις πλευρές των εργασιών ταυτοποίησης ή τουλάχιστον να δίνει στο χρήστη την ικανότητα να εξετάζει τις ενδιαφέρουσες πλευρές, όπως για παράδειγμα οι κατασκευαστές να μπορούν να εξετάζουν τα νέα χαρακτηριστικά των προϊόντων τα οποία κατασκευάζουν ενάντια στους ανταγωνιστές ή να μπορούν να αναγνωρίζουν υπάρχων περιορισμούς οι οποίοι μπορούν να λειτουργήσουν ως πιθανές κατευθύνσεις έρευνας[IRV13]. Οι περισσότερες τεχνικές ταυτοποίησης είναι προσαρμοσμένες για δικούς τους συγκεκριμένους στόχους.

2.2 Συστήματα Ταυτοποίησης

Έχουν δημιουργηθεί πολλά συστήματα που μπορούν να παράγουν σύνολα δεδομένων που αναφέρονται στα ίδια αντικείμενα με σκοπό την αξιολόγηση αλγορίθμων ταύτισης αντικειμένων. Τα σημεία αναφοράς μετρήσεων σχετικά με σχήμα, οντολογία και ταυτοποίηση αντικειμένων είναι ενδιαφέρον μιας και η σημασιολογία η οποία εμπλέκεται στη διαδικασία ίσως προκαλέσει έναν αριθμό θεμάτων σχετικά με το ποιος είναι ο ακριβής σωστός αριθμός πληροφοριών. Τα δεδομένα, βάσει των οποίων λειτουργούν τα συστήματα ταυτοποίησης, και λόγω της ετερογένειας της οποίας έχουν, ίσως να μην αναφέρονται στο ίδιο θέμα. Επομένως είναι σημαντικό να υπάρχει η ικανότητα να παράγονται δεδομένα τα οποία περιέχουν παρόμοια αντικείμενα. Η ταυτοποίηση αντικειμένων μπορεί να χρησιμοποιηθεί κατά την πάροδο του χρόνου ενσωμάτωσης πληροφοριών στον ιστό ή για απάντηση ερώτησης όσο αφορά ένα

αντικείμενο. Επομένως το μέγεθος των πληροφοριών θα επηρεάσει αρκετά την εκτέλεση εργασίας ταυτοποίησης αντικειμένων. Για τον λόγο αυτό τα εργαλεία ταυτοποίησης με πληροφορίες διαφορετικών μεγεθών είναι σημαντικά για να κατανοηθεί πως αυξάνονται σε σχέση με τον χρόνο. Υπάρχουν συστήματα τα οποία έχουν τη δυνατότητα να ελέγχουν το μέγεθος των δεδομένων τα οποία παράγονται [IRV13].

Τα διάφορα συστήματα και εργαλεία που υπάρχουν δεν μπορούν να συγκριθούν μεταξύ τους ούτε μπορεί να ειπωθεί ότι ένα από αυτά είναι το καλύτερο και υπερτερεί των άλλων. Κατά την συγγραφή πολλών ερευνών αρκετοί συγγραφείς έχουν παρουσιάσει συστήματα ταυτοποίησης με διαφορετικές ρυθμίσεις και δυνατότητες. Για παράδειγμα, όλα παρέχουν δυνατότητες και υπηρεσίες αναζήτησης αντικειμένων και στοχεύουν στο να μη είναι χρονοβόρα η διαδικασία αφαίρεσης διπλοτύπων σε μεγάλες ημι-δομημένες λίστες. Για τους σκοπούς αυτής της μεταπτυχιακής διατριβής θα χρησιμοποιηθεί το σύστημα EMBench το οποίο είναι ένα σωστό σύστημα αξιολόγησης συστημάτων ταυτοποίησης αντικειμένων[IRV13],[IV14]. Δέχεται ως δεδομένα ένα σύνολο διαμόρφωσης παραμέτρων και παράγει μια σειρά περιπτώσεων δοκιμής για αξιολόγηση ενός συστήματος ταυτοποίησης αντικειμένων. Αποτελείται από τρία κύρια μέρη: (i) ένα χώρο από δεδομένα που θα χρησιμοποιηθούν στην κατασκευή των συλλογών ταυτοποίησης των σεναρίων, παράλληλα με τα στοιχεία (που ονομάζονται διαχωριστές) που καταλαμβάνουν αυτό το χώρο από διάφορες πηγές που είναι σημαντικές και ποιοτικές (ii) μια μηχανή παραγωγής ταυτοποίησης που συνθέτει τα δεδομένα στους χώρους δεδομένων για να διατυπώσει μια συλλογή από ταυτοποιήσεις και (iii) μια σειρά από τροποποιητές που τροποποιούν με διάφορους τρόπους τα δεδομένα στη συλλογή ταυτοποίησης και κατασκευάζουν μια νέα συλλογή ταυτοποίησης με υψηλό βαθμό ετερογένειας. Για να διευκολύνει αυτές τις λειτουργίες, το EMBench είναι γενικά πλήρως παραμετροποιημένο μέσω ενός αρχείου διαμόρφωσης[IRV13].Επίσης μέσω του τεμαχιστή του συστήματος υπάρχει η δυνατότητα πολλών τροποποιήσεων όπως ποιες ιδιότητες πινάκων θα συλλεχθούν, ποιος θα είναι ο μέγιστος αριθμός αντικειμένων, πόσα αντικείμενα θα επιλεχθούν , ποιος τροποποιητής θα χρησιμοποιηθεί και σε ποιο ποσοστό.Το EMBench χρησιμοποιεί εκτεταμένο datalog με λειτουργία συνόλου και ομαδοποίησης [IRV13].Επιπλέον έχει περισσότερη εκφραστική δύναμη και περισσότερη ελαστικότητα στην προδιαγραφή των δοκιμαζόμενων πληροφοριών.

Ένα άλλο σύστημα που σχεδιάστηκε από τους Sarawagi et al. είναι το ALIAS [SB02], [SBK+02], [SK03] το οποίο επιτρέπει την αυτόματη δόμηση της λειτουργίας deduplication με το να χρησιμοποιεί μια καινούργια μέθοδο: να ανακαλύπτει την αλληλεπίδραση ενδιαφέρων ζευγαριών εκπαίδευσης. Στόχος του ALIAS είναι να αυτοματοποιηθεί η χειροκίνητη και χρονοβόρα διαδικασία αφαίρεσης διπλοτύπων σε μεγάλες ημι-δομημένες λίστες, να δημιουργήσουν διάφορες εφεδρικές λειτουργίες και να εκμεταλλευτούν τις διαφωνίες ανάμεσα τους για να ανακαλύψουν νέα είδη ασυνεπειών ανάμεσα στα διπλότυπα στη βάση δεδομένων. Το σύστημα ALIAS βασίζεται στην μάθηση η οποία μέθοδος αυτή θα ήταν απίστευτη ακριβής όταν οι αριθμοί των αρχείων είναι μεγάλοι. Επίσης χειρίζεται την λειτουργία deduplication ως ένα μαύρο κουτί για πάρει ένα άμεσο προϊόν από τις καταχωρήσεις. Οι συγγραφείς ασχολήθηκαν κυρίως με βάσης δεδομένων που περιείχαν δημοσιεύσεις οπόταν δεν υπάρχουν αποτελέσματα για άλλου είδους πινάκων στηλών που να περιέχουν διαφορετικές ιδιότητες. Οι Sarawagi και Bhamidipaty επέλεξαν κατηγοροποιητή ώστε να είναι ευέλικτος, ακριβείς και να μπορεί να εκπαιδευτεί εύκολα και χρησιμοποίησαν την μέθοδο “decision tree” για την κατηγοριοποίηση των συνόλων δεδομένων. Επίσης στο σύστημα τους ,μπορεί να επιλεχθεί λειτουργία για την ταυτοποίηση , χωρίς όμως να υπάρχουν αρκετές επιλογές. Μπορεί όμως να προστεθεί από τον χρήστη μια άλλη λειτουργία με προγραμματισμό C++. Δεν υπάρχει όμως η δυνατότητα επιλογής ποσοστού για τον κατηγοροποιητή.

Οι Dong et al. παρουσίασαν το Personal Information Management (PIM) το οποίο υποστηρίζει την αναζήτηση πληροφοριών στον υπολογιστή σε ένα καλύτερο επίπεδο . [DHM05]. Έδωσαν έμφαση στη δημιουργία ενός συστήματος που να μπορεί να εξετάζει πληροφορίες από μια ποικιλία πηγών στον υπολογιστή, για παράδειγμα ηλεκτρονική διεύθυνση, επαφές κ.α. για να εξάγει τεκμήρια πολλαπλών κατηγοριών: πρόσωπο/άτομο, μήνυμα, κ.α. Επιπλέον εξάγει συνδέσεις μεταξύ των τεκμηρίων όπως senderOf, earlyVersionOf, authorOf και publishedIn τα οποία παρέχουν την βάση της αναζήτησης. Ωστόσο, αυτές οι πηγές πληροφοριών είναι ετερογενής και διαρκούν πολλά χρόνια ενώ ένα πραγματικό αντικείμενο τυπικά αναφέρεται με πολλούς διαφορετικούς τρόπους. Το σύστημα τους λειτουργεί καλύτερα σε περιπτώσεις των τεχνικών σύγκλισης βασικών αναφορών.

Άλλο σύστημα που χρησιμοποιήθηκε από τους Shen et al. είναι το Source Conscious Compiler for Entity Resolution (Soccer) [SDV+07]. Το σύστημα δημιουργήθηκε για να

λύσει το πρόβλημα που εμφανίζεται σε πολλές εφαρμογές οι οποίες ενσωματώνουν πληροφορίες από πολλές πηγές. Το Soccer έχει ως στόχο να ορίσει το πρόβλημα ταυτοποίησης. Συγκεκριμένα κάθε αλγόριθμος ταυτοποίησης χειρίζεται ως ένας χειριστής μαύρου κουτιού (blackbox operator) ο οποίος ονομάζεται matcher .Αν και υποστηρίζει αρκετές βάσεις δεδομένων εντούτοις εφαρμόζεται κάθε φορά για την ταυτοποίηση αναφορών μόνο ενός τύπου.

Οι Miklos et al. [MBB+10] δημιούργησαν ένα άλλο σύστημα, το Entity Name System (ENS), το οποίο παρέχει δωρεάν, ανοικτή υπηρεσία στους χρήστες του διχτύου έτσι ώστε να μπορούν να σχολιάζουν το περιεχόμενο του δικτύου με αναφορές στα αντικείμενα. Η κύρια λειτουργία του ENS είναι να επεξεργάζεται τις αναζητήσεις αντικειμένων και να επιστρέφει μοναδικό αναγνωριστή του αντικειμένου. Το σύστημα τους είναι προσαρμοσμένο στο προσανατολισμό του αντικειμένου και όχι σε μοντέλο που βασίζεται σε λέξεις κλειδιά. Έχει αρκετές δυνατότητες όπως το να μπορεί να αποθηκεύει μεγάλο αριθμό περιγραφών αντικειμένων, να προσδιορίζει τους μοναδικούς αναγνωριστές στις περιγραφές και υποστηρίζει ανοικτές διαθέσιμες βάσης δεδομένων. Υστερεί όμως σε επιλογές τροποποιητών, μιας και στο σύστημα υπάρχουν αυτοματοποιημένες επιλογές αλλά για να τρέξουν κάποιες που δεν υπάρχουν αυτές θα πρέπει να προστεθούν χειροκίνητα στο σύστημα.

Ακόμη ένα σύστημα ταυτοποίησης αντικειμένων είναι το Semantic Web INstance Generation (SWING) [FMN+11] το οποίο αποτελείται από μια βάση εξέτασης υποθέσεων, που κάθε μια απ αυτές αντιπροσωπεύεται από μια βάση ομάδων/παραδειγμάτων και σχετικών δηλώσεων που έχουν δημιουργηθεί από μια αρχική βάση δεδομένων πραγματικών συνδεδεμένων δεδομένων τα οποία λήφθηκαν από το διαδίκτυο . Το σύστημα SWING παράγει πληροφορίες για τη συγκριτική αξιολόγηση συστημάτων ταυτοποίησης αντικειμένων το οποίο εστιάζει στο να παρέχει υποδομή για αξιολόγηση σημασιολογικών τεχνολογιών και έχει χρησιμοποιηθεί για την αξιολόγηση τεχνικών σε πρόσφατες εκστρατείες της OAEI (Ontology Alignment Evaluation Initiative) [IRV13]. Η επίτευξη πληροφοριών βασίζεται στην ανάκτηση πληροφοριών από ήδη υπάρχων συνδεδεμένων πηγών πληροφοριών και έπειτα αλλάζονται για να λειτουργήσουν ως ανάλυση δεδομένων. Επίσης βασίζεται στο DL (Description Logics) επομένως χρησιμοποιείται η εκφραστική δύναμη της προδιαγραφής του μετασχηματισμού αλλά δεν παρέχει κανένα έλεγχο στην διανομή των ταξινομημένων αξιών[FMN+11]. Οι διανομές των τιμών σε μια ομάδα τυπικά

ακολουθεί αυτές των πηγών και η διανομή στα μετασχηματισμένα δεδομένα θα εξαρτηθεί μόνο στην αρχική διανομή και στην αλλαγή που έχουν εφαρμοστεί. Επιπρόσθετα, προσφέρει την ικανότητα να δημιουργήσει βάση δεδομένων διαφορετικών μεγεθών. Υπάρχουν αρκετές δυνατότητες στο εν λόγω σύστημα και αρκετές επιλογές διάφορων κατηγοριών τροποποιητών. Υποστηρίζει συνδεδεμένες πηγές συνόλων δεδομένων αλλά δεν υπάρχει η δυνατότητα ο χρήστης να μπορεί να καθορίσει τον αριθμό και τις ιδιότητες των αντικειμένων[IRV13].

Τα περισσότερα προαναφερθέντα συστήματα των διαφόρων ερευνητών δημιουργούν δεδομένα τα οποία έχουν χαμηλό αριθμό παρατηρήσεων και όχι πολλές ιδιότητες στην ίδια ομάδα (collection) και επιπλέον ούτε σημαντικά λάθη που αυτό έχει ως συνέπεια να μη συλλέγονται πιθανώς όλες οι πληροφορίες που αναφέρονται στο ίδιο αντικείμενο.

2.3 Τεχνικές Ταυτοποίησης

Υπάρχουν διάφορες προσεγγίσεις που έγιναν από συγγραφείς για την επίλυση του προβλήματος ταυτοποίησης αντικειμένων. Σε μια προσέγγιση που προτάθηκε μπορεί απλά να διαιρεθούν αυτές οι προσεγγίσεις σε δύο κατηγορίες με βάση τις ιδιότητες που θεωρούνται ότι μπορούν να επιλύσουν αυτό το πρόβλημα[MSM16]. Η πρώτη λαμβάνει υπόψη μόνο μη διαχρονικά χαρακτηριστικά δηλαδή χαρακτηριστικά τα οποία δεν αλλάζουν κατά την πάροδο του χρόνου όπως είναι το όνομα. Η δεύτερη λαμβάνει υπόψη και διαχρονικά καθώς και μη-διαχρονικά γνωρίσματα. Τα διαχρονικά αναφέρονται σε αυτά που μπορεί να αλλάξουν όπως είναι το τηλέφωνο, η διεύθυνση ή το επίθετο. Στο πλαίσιο της πρώτης κατηγορίας εμπίπτουν πολλές τεχνικές. Στην κατηγορία αυτή, οι σχέσεις μεταξύ των παρατηρήσεων συστάθηκαν με τη χρήση διαφορετικών στρατηγικών που βασίζονται σε κανόνες ή μηχανικής μάθησης με βάση στρατηγικής. Στη δεύτερη κατηγορία χρησιμοποιήθηκαν τα διαχρονικά χαρακτηριστικά.

Κατή την διάρκεια της έρευνας που έκαναν, οι Yin et al. συμπερίλαβαν αρκετές τεχνικές και αναφέρθηκαν σε αυτές ξεχωριστά για την κάθε μια. Μια από τις τεχνικές που πρότειναν ήταν το DISTINCT[YHY07]. Σύμφωνα με την τεχνική αυτή, ο SVM(Support Vector Machines) ταξινομητής χρησιμοποιείται για τον προσδιορισμό διαφορετικών τύπων συνδέσεων. Για το σύνολο εκπαίδευσης, χρησιμοποιήθηκε μια σειρά από

διακριτά αντικείμενα και ο συγγραφέας με μια εγγραφή η οποία δεν έχει σχέση με μια άλλη εγγραφή με συν-συγγραφείς ή συνέδρια δεν λήφθηκε υπόψη. Στόχος τους ήταν μέσω του συστήματος DISTINCT να διακριθούν και μετέπειτα να συγχωνευτούν οι ταυτότητες αντικειμένων που αναφέρονται στον ίδιο συγγραφέα. Θεώρησαν ότι μια σειρά δεδομένων μοιάζουν αν περιέχουν κείμενο που ταυτίζεται πχ. αναφορές σε συγγραφείς με ονόματα που ταυτίζονται και δύο αναφορές είναι ισοδύναμες αν αναφέρονται στο ίδιο αντικείμενο. Στόχος τους ήταν να ομαδοποιήσουν αυτές τις σειρές ώστε να περιλαμβάνουν παρόμοιες σειρές. Μια άλλη προσέγγιση που προτάθηκε από τους Tang et al. [TZY+08], ήταν να ταυτοποιήσουν άρθρα με κατηγοριοποιητές το όνομα του ερευνητή και του συγγραφέα δημοσίευσης, θέτοντας σχέσεις μεταξύ των ιδιοτήτων που χρησιμοποίησαν. Οι σχέσεις που έθεσαν ήταν τύπου , δύο άρθρα έχουν τον ίδιο δευτερεύον συγγραφέα, δύο άρθρα εκδόθηκαν στην ίδια τοποθεσία κτλ. Στόχος ήταν να μπορεί να προσδιοριστεί ότι δύο συγκεκριμένα άρθρα πρέπει να ανατεθούν στο ίδιο πρόσωπο.

Οι μέθοδοι της δεύτερης κατηγορίας λαμβάνουν υπόψη τα διαχρονικά χαρακτηριστικά και τα μη διαχρονικά χαρακτηριστικά. Πρώτα όμως λαμβάνονται υπόψη τα διαχρονικά χαρακτηριστικά. Οι εγγραφές κατατάσσονται, κατά την διαδικασία ταυτοποίησης σε χρονική σειρά. Τα χαρακτηριστικά που λαμβάνονται υπόψη είναι η σύνδεση, ο συν-συγγραφέας και ο χρόνος. Επίσης προτάθηκε μια νέα τεχνική ταυτοποίησης αντικειμένων που ονομάζεται EMTB[MSM16] και η οποία λαμβάνει υπόψη τα χαρακτηριστικά, όπως συν-συγγραφέας, σύνδεση, ηλεκτρονική διεύθυνση-ταυτότητα συγγραφέα, τόπος , τις αναφορές και το χρόνο δημοσίευσης.

Όπως ανάφεραν οι Mishra et al. [MSM16] το αποτέλεσμα που προκύπτει από οποιαδήποτε από τις τεχνικές ταυτοποίησης αντικειμένων πρέπει να συγκριθεί με το πρότυπο αποτελεσμάτων για τη μέτρηση της αποτελεσματικότητας της τεχνικής. Αλλά η εξασφάλιση αυτού του πρότυπου είναι δύσκολη όπως ανάφεραν χαρακτηριστικά. Υπάρχουν κάποιες εργασίες στο πρόσφατο παρελθόν, οι οποίες χρησιμοποιούνται για να κρίνουν την απόδοση ενός αλγορίθμου που αναπτύχθηκε για την ταυτοποίηση αντικειμένων σε βιβλιογραφική βάση δεδομένων.

Οι Herziq et al. μέσω τον συστημάτων KW και QR έτρεξαν πειράματα για την ταυτοποίηση αντικειμένων[HT12]. Σκοπός τους, ήταν κάθε φορά να χρησιμοποιείται μια βάση δεδομένων ως η πηγή δεδομένων και οι υπόλοιπες δύο ως τα σύνολα δεδομένα των στόχων. Στόχος τους ήταν κάθε φορά ένα αντικείμενο που υπάρχει στην

βάση δεδομένων που ορίστηκε ως πηγή, να ταυτοποιείται με τα ίδια αντικείμενα αλλά που έχουν διαφορετικές εγγραφές που υπάρχουν στις άλλες δύο βάσεις που ορίστηκαν ως στόχος.

Οι Ioannou et al. [IRV13] με την σειρά τους χώρισαν τις τεχνικές ταυτοποίησης αντικειμένων σε πέντε κύριες κατηγορίες. Η πρώτη κατηγορία αποτελείται από τις προσεγγίσεις που χρησιμοποιούν τεχνικές ομοιότητας σε ατομικό επίπεδο για να αποφασιστεί πόσο κοντά είναι δύο δομές. Χειρίζονται περιπτώσεις όπως "John D. Smith" έναντι "J. D. Smith ", και " Transactions on Knowledge and Data Engineering " έναντι " IEEE Trans. Knowl. Data Eng.". Η ταυτοποίηση γίνεται με την ανίχνευση της ομοιότητα μεταξύ των τιμών του κειμένου που βρέθηκαν στα αντικείμενα.

Η δεύτερη κατηγορία περιλαμβάνει τις τεχνικές που χειρίζονται τα αντικείμενα ως ένα σύνολο ατομικών τιμών. Ένα κλασσικό παράδειγμα είναι μια σχέση εγγραφής που αντιπροσωπεύει ένα αντικείμενο. Για να μειωθεί το πρόβλημα και να γίνει ένα το οποίο θα συγκρίνει ατομικές τιμές, ένωσαν όλες τις ατομικές τιμές για κάθε αντικείμενο σε ένα στοιχειοσύνολο το οποίο στη συνέχεια χρησιμοποιήθηκε σε συγκρίσεις ατομικών τιμών ως αντιπρόσωπος του αντικειμένου.

Η τρίτη κατηγορία εστιάζει στην συλλογική ταυτοποίηση. Αντί να χρησιμοποιεί τις πληροφορίες από δύο αντικείμενα μόνο, η συλλογική ταυτοποίηση προτείνει την αξιοποίηση των πληροφοριών σχετικά με τα ταχτοποιημένα αντικείμενα μεταξύ δύο συνόλων. Για παράδειγμα οι συγγραφείς στις δημοσιεύσεις. Γνωρίζοντας ότι μια δημοσίευση έχει ως συγγραφείς τον α, β και τον γ και μια άλλη δημοσίευση έχει ως συγγραφείς τον β’ και τον γ μπορεί να πει κανείς με μεγάλη σιγουριά ότι το β περιγράφει τον ίδιο συγγραφέα με τον β’ .

Στην τέταρτη κατηγορία υπάρχουν οι μέθοδοι που βασίζονται στον αποκλεισμό. Η ιδέα του αποκλεισμού είναι ότι αντί να συγκρίνει κάθε αντικείμενο με όλα τα άλλα αντικείμενα, τα αντικείμενα χωρίζονται σε ομάδες και κάθε αντικείμενο συγκρίνεται μόνο με τα αντικείμενα που υπάρχουν στην ίδια ομάδα. Η πρόκληση είναι να δημιουργηθούν ομάδες από αντικείμενα που είναι πιθανό να αναφέρονται στο ίδιο πράγμα του πραγματικού κόσμου. Πολλές τεχνικές συνδέουν το κάθε αντικείμενο με μια τιμή , συνοψίζοντας τις τιμές των επιλεγμένων χαρακτηριστικών και στη συνέχεια , λειτουργούν αποκλειστικά σε αυτό.

Η τελευταία κατηγορία περιλαμβάνει τις τεχνικές που εκμεταλλεύονται τις πληροφορίες του διαγράμματος. Στην ταυτοποίηση διαγράμματος , στόχος είναι να προσδιοριστούν οι αντιστοιχίες μεταξύ των χαρακτηριστικών ή των δομών του διαγράμματος που επιδεικνύουν την ίδια έννοια.

Κεφάλαιο 3

Δεδομένα Ταυτοποίησης Αντικειμένων

Το κεφάλαιο μελετά τις ιδιότητες καθώς και τις οικογένειες των συνόλων δεδομένων που συνήθως χρησιμοποιούνται στην ερευνητική κοινότητα για την αξιολόγηση αλγορίθμων ταύτισης αντικειμένων. Χωρίζεται σε τρείς ενότητες οι οποίες είναι: υφιστάμενες μελέτες, ιδιότητες και καθορισμός οικογενειών και συνόλων. Μέσα από υφιστάμενες μελέτες θα γίνει αναφορά σε ερευνητές που έχουν ασχοληθεί με το πρόβλημα της ταυτοποίησης αντικειμένων και Θα αναφερθούν οικογένειες συνόλων δεδομένων που συνήθως χρησιμοποιήσανε. Στο κεφάλαιο θα παρουσιαστούν συγκεντρωτικά οι ιδιότητες που χρησιμοποιούν συνήθως οι ερευνητές. Τέλος θα καθοριστούν οι οικογένειες που θα δημιουργηθούν σε αυτή τη μεταπτυχιακή διατριβή βάση των όσων συλλέχθηκαν στο κεφάλαιο, ως επίσης και οι ιδιότητες της κάθε οικογένειας έχοντας υπόψη το τι χρησιμοποιούν συνήθως οι ερευνητές.

3.1 Υφιστάμενα Σύνολα Δεδομένων

Στην εν λόγω ενότητα γίνεται περιγραφή των υφιστάμενων συστημάτων συνόλων δεδομένων που έχουν ήδη χρησιμοποιηθεί. Τα συστήματα αυτά έχουν αρκετούς περιορισμούς και μέσω της καινούριας δημιουργίας συνόλων ίσως να υπάρξει βελτίωση στη λειτουργία τους. Οι περιορισμοί που υπάρχουν αναφέρονται στον μη συνδυασμό διαφόρων επιπέδων τύπων, πολυπλοκότητας και κλιμάτων. Ως αποτέλεσμα δεν δύναται μια πλήρη και ακριβή αξιολόγηση ενός αλγορίθμου ταύτισης αντικειμένων. Επιπλέον ο χρήστης δεν είχε τη δυνατότητα και ευελιξία πολλών τροποποιήσεων και χαρακτηριστικών κατά τη δημιουργία των συνόλων δεδομένων και ούτε μπορούσε να ελέγχει τη δομή και την μορφή των συνόλων που είχαν δημιουργηθεί.

Μια κοινή προσέγγιση που ακολουθείτε από πολλές τεχνικές ταυτοποίησης προκειμένου να αξιολογηθούν οι τεχνικές ταυτοποίησης αντικειμένων, είναι να χρησιμοποιηθεί κάποιο δοκιμαστικό σύνολο δεδομένων και να μετρηθεί η επιτυχία του συστήματος τους σε αυτό. Η δημιουργία συνόλων από δεδομένα που περιλαμβάνουν δημοσιεύσεις σε συνδυασμό με το όνομα του συγγραφέα, χρονολογία δημοσίευσης, τίτλος του άρθρου και τίτλος του περιοδικού που δημοσιεύτηκε παρατηρείτε συχνά σε βιβλιογραφικές αναφορές[IRV13],[YHY07]. Ακόμη ένα σύνολο δεδομένων που χρησιμοποιείται συχνά είναι οι ταινίες, μαζί με το όνομα του ηθοποιού, χρονολογία της ταινίας και τίτλος ταινίας. Τέλος συχνά στην ερευνητική κοινότητα χρησιμοποιηθήκαν και σύνολα δεδομένα που αναφέρονται σε ακαδημαϊκούς και πανεπιστήμια. Οι πιο πολλοί ερευνητές/συγγραφείς δημιουργούν οικογένειες αναλόγως των ιδιοτήτων που θέλουν να συμπεριλάβουν, που να περιλαμβάνουν τα πιο πάνω αλλά με διαφορετικές παραμέτρους ώστε να μπορεί να αξιολογηθεί ο αλγόριθμος σε διαφορετικά σύνολα από δεδομένα αλλά να διαφέρουν σε κάποιες παραμέτρους όπως αριθμό των αντικειμένων που αποτελούνται, διαφορετικών αριθμό ιδιοτήτων, διαφορετικό τροποποιητή, διαφορετικό ποσοστό τροποποιητή. Το Cora είναι ένα από αυτά τα δοκιμαστικά σύνολα δεδομένων και χρησιμοποιήθηκε για την αξιολόγηση των διαφόρων αλγορίθμων. Το σύνολο δεδομένων περιέχει 1.295 επιστημονικές δημοσιεύσεις από την Cora Computer Science Research Paper Engine , και πιο συγκεκριμένα περιλαμβάνει περίπου 9.700 περιγραφές για 2.800 εγγραφές συγγραφέων του πραγματικού κόσμου . Η ακρίβεια συνήθως μετράται σε όρους ακρίβειας και ανάκλησης που βασίζονται στις ταυτοποιήσεις που δίνονται μεταξύ των διαφορετικών περιγραφών των συγγραφέων [IRV13]. Επίσης πολλοί συγγραφείς στην ερευνητική κοινότητα έχουν χρησιμοποίησαν τα σύνολα δεδομένων Amazon[1], IMDb[2], DBLP[3], DBpedia[4]. Το σύνολο δεδομένων DBpedia είναι μια δοδημένη αναπαράσταση της Wikipedia, η οποία περιέχει περισσότερα από εννέα εκατομμύρια διαφόρων τύπων αντικειμένων[HT12]. Το σύνολο δεδομένων DBLP περιέχει βιβλιογραφία σχετικά με την επιστήμη των υπολογιστών και περιλαμβάνει περισσότερα από 3,3 εκατομμύρια δημοσιεύσεις[YHY07]. Το σύνολο δεδομένων IMDb περιέχει πληροφορίες σχετικά με ταινίες και συμπεριλαμβάνει περίπου 859.000 αντικείμενα ενώ το σύνολο δεδομένων Amazon περιέχει πληροφορίες σχετικά με DVDS, και VHS videos και αποτελείται από περίπου 115.000

[1] http://www.amazon.com/
[2] http://www.imdb.com/
[3] http://dblp.uni-trier.de/faq/What+is+dblp.html
[4] http://dbpedia.org/About

αντικείμενα[HT12]. Σύνολα δεδομένων που χρησιμοποιήθηκαν σε πολλές έρευνες είναι και ηλεκτρονικές βιβλιοθήκες βιβλιογραφίας όπως την ACM Digital Library και Citeseer.

3.2 Υφιστάμενες Μελέτες για τη Δημιουργία Συνόλων

Στην υποενότητα που ακολουθεί γίνεται εκτενής αναφορά στο τρόπο δημιουργίας συνόλων από προηγούμενες μελέτες. Στις μελέτες αυτές αναφέρεται ο αριθμός των ιδιοτήτων που έχουν χρησιμοποιηθεί αλλά και ο αριθμός δείγματος. Ένα σημαντικό στοιχείο που παρατηρείται είναι πως τις περισσότερες φορές δεν καθορίζεται επαρκής ετερογένεια και συντακτική παραλλαγή.

Για να καταστεί δυνατόν ένας καινοτόμος καθορισμός αλλά και η δημιουργία καινούργιων οικογενειών συνόλων δεδομένων πρέπει να γίνει αναφορά σε σχετικές έρευνες ταυτοποίησης ώστε να δημιουργηθούν σύνολα δεδομένων με διαφορετικές παραμέτρους με αυτά που έχουν χρησιμοποιηθεί μέχρι τώρα στην ερευνητική κοινότητα με σκοπό να γίνεται καλύτερα η ταυτοποίηση. Οι ιδιότητες οι οποίες θα αποτελούν και τους πίνακες στηλών που θα περιέχει κάθε οικογένεια είναι σημαντικό να επιλεχθούν σωστά ώστε να σχετίζονται μεταξύ τους και να αναφέρονται στο ίδιο πράγμα. Για να γίνει αυτό θα πρέπει πρώτα να ερευνηθεί ποιες ιδιότητες και οικογένειες χρησιμοποιούνται συνήθως. Οι ποιο συνηθισμένες οικογένειες που χρησιμοποιούνται συνήθως είναι αυτές που αναφέρονται σε δημοσιεύσεις, ταινίες και ερευνητές. Πολλοί ερευνητές χρησιμοποίησαν είτε δοκιμαστικά σύνολα δεδομένων ή δημιούργησαν δικά τους από πηγές συνόλων δεδομένων που περιλάμβαναν ιδιότητες όπως όνομα, τίτλος, τοποθεσία, χρονολογία, περιοδικό και άλλες.

Οι Yin et al. χρησιμοποίησαν στο πείραμα τους δημοσιεύσεις με ιδιότητες το όνομα του συγγραφέα, τον τίτλο, την χρονολογία, το συνέδριο, την τοποθεσία και τον εκδότη[YHY07]. Στόχος τους ήταν μέσω του συστήματος Distinct να διακριθούν και μετέπειτα να συγχωνευτούν οι ταυτότητες αντικειμένων που αναφέρονται στον ίδιο συγγραφέα. Στη μελέτη τους χρησιμοποίησαν ως αρκετά μεγάλο αριθμών δειγμάτων γύρω στις 127000 και ως παράμετρο μόνο την ιδιότητα του ονόματος του συγγραφέα και δεν συμπεριέλαβαν συγγραφείς που δεν έχουν περισσότερο από δύο άρθρα. Οι

γειτονικές πλειάδες μιας αναφοράς είναι οι πλειάδες που μπορούν να συνδεθούν. Η σημασιολογική σημασία των γειτονικών πλειάδων καθορίζεται από τη διαδρομή σύνδεσης και αυτό το έπραξαν με το πεδίο άρθρο-κλειδί που χρησιμοποίησαν στους πίνακες στήλες. Θεώρησαν ότι μια σειρά παραπομπών μοιάζουν αν περιέχουν κείμενο που ταυτίζονται πχ. αναφορές σε συγγραφείς με ονόματα που ταυτίζονται και δύο αναφορές είναι ισοδύναμες αν αναφέρονται στο ίδιο αντικείμενο. Στόχος τους ήταν να ομαδοποιήσουν αυτές τις σειρές ώστε να περιλαμβάνουν παρόμοιες σειρές.

Στην δική τους έρευνα οι Ioannou et al. [IRV13] χρησιμοποίησαν για τις αξιολογήσεις αλγορίθμων σύνολα από δεδομένα με περισσότερες ιδιότητες σε δημοσιεύσεις όπως το τίτλο του άρθρου, το όνομα του συγγραφέα, τη χρονολογία και την τοποθεσία. Χρησιμοποίησαν επίσης δεδομένα που αναφέρονταν σε ακαδημαϊκούς, με το όνομα αυτού και το πανεπιστήμιο. Χρησιμοποίησαν διαφορετικό αριθμό αντικειμένων στα σύνολα δεδομένων ώστε να αξιολογήσουν τον χρόνο που θα χρειαστεί ο αλγόριθμος για την ταυτοποίηση. Στη περίπτωση αυτή υπήρχε πλεονέκτημα ο αριθμός των ιδιοτήτων αλλά ο αριθμός των δειγμάτων ήταν στα 10000, αρκετά μικρότερος σε σχέση με τους Yin et al. [YHY07]. Παρόμοια σύνολα δεδομένων χρησιμοποίησαν οι Mishra et al. [MSM16]. Χρησιμοποίησαν οχτώ σύνολα δεδομένα από τα οποία το καθένα αντιστοιχεί σε κατάλογο όλων των εγγράφων που δημοσιεύονται από τον συγγραφέα με ένα όνομα. Χρησιμοποίησαν πληροφορίες χαρακτηριστικών όπως συγγραφέα, συν-συγγραφέα, ηλεκτρονική διεύθυνση και το φορέα στον όποιο ανήκουν αλλά είχαν ως περιορισμό το μικρό αριθμό δειγμάτων που χρησιμοποίησαν. Οι Gu et al. [GB06] έθεσαν σαν συγκρίσιμες μεταβλητές το όνομα, το επίθετο, την ηλικία και άλλες τις οποίες άντλησαν από διάφορους πίνακες συχνοτήτων όπου και σε αυτή τη περίπτωση ο αριθμός δείγματος ήταν πάρα πολύ μικρός. Χρησιμοποίησαν το εργαλείο παραγωγής δεδομένων από σύνολα δεδομένων για την δημιουργία δοκιμαστικών συνόλων για το πείραμα που έκαναν. Δημιούργησαν αντίγραφα των αρχικών εγγραφών αλλά με διπλότυπες εγγραφές. Τα δεδομένα που δημιουργήθηκαν τροποποιούνταν με διάφορες παραμέτρους με τυχαία επιλογή αυτών. Ο βαθμός των τροποποιήσεων καθορίστηκε από τις πιθανές αντίστοιχες πιθανότητες. Αυτή η τυχαία επιλογή ίσως να δημιουργεί πρόβλημα στη χρήση του σωστού αλγόριθμου για τη ταυτοποίηση των αντικειμένων.

Στο δικό τους πείραμα οι Herziq et al. [HT12] , χρησιμοποίησαν τα σύνολα δεδομένων Amazon, IMdb, DBpedia, για να ταυτοποιήσεων εγγραφές αντικειμένων που αναφέρονταν στο ίδιο αντικείμενο. Χρησιμοποίησαν τις ιδιότητες ταινία, όνομα

σκηνοθέτη και ηθοποιού , χρονολογία παραγωγής την ταινίας και τον τύπο της ταινίας. Σκοπός ήταν κάθε φορά να χρησιμοποιείται μια βάση δεδομένων ως πηγή δεδομένων και οι υπόλοιπες δύο ως τα σύνολα δεδομένα των στόχων. Οι Herziq et al. χρησιμοποίησαν πάρα πολύ μεγάλο αριθμό δείγματος, ένα εκατομμύριο, όπου στη περίπτωση αυτή δημιουργούνται θέματα στην αξιολόγηση της επαναληψιμότητα και την απευθείας σύγκριση μεταξύ των ταυτοποιημένων αντικειμένων. Παρόμοια μέθοδο ακολούθησαν οι Kopche et al. [KTR10], χρησιμοποιώντας όμως ως ιδιότητες από δημοσιεύσεις δηλαδή, τίτλος, όνομα συγγραφέα , τοποθεσία και χρονολογία. Επίσης δημιούργησαν ακόμη ένα σύνολο δεδομένων που περιλάμβανε ιδιότητες όπως όνομα, περιγραφή, κατασκευαστής και τιμή. Χρησιμοποίησαν το σύστημα FEVER για την αξιολόγηση αλγορίθμων σχετικά με την ταυτοποίηση διαφορετικών εγγραφών που αναφέρονταν και ίδιο αντικείμενο. Οι Dong et al. χρησιμοποίησαν ως ιδιότητες το όνομα, ηλεκτρονικό ταχυδρομείο, συν-γραφέας και ηλεκτρονικό ταχυδρομείο [DHM05]. Αυτές οι ιδιότητες χρησιμοποιηθήκαν για το σύνολο δεδομένο πρόσωπό. Δημιούργησαν ακόμη τρία σύνολα δεδομένων που ονόμασαν άρθρο, συνέδριο και περιοδικό. Το σύνολο δεδομένων άρθρο αποτελείτε από ιδιότητες όπως τίτλος, χρονολογία, σελίδες, συγγραφέας, και που δημοσιεύτηκε. Για το σύνολο δεδομένων συνέδριο επέλεξαν ως ιδιότητες το όνομα, την χρονολογία και την τοποθεσία και τέλος για το σύνολο δεδομένων περιοδικό επέλεξαν τις ιδιότητες όνομα, χρονολογία, τόμος και αριθμός. Χρησιμοποιώντας το σύστημα PIM, δημιούργησαν καινούργια δεδομένα όπου περιλαμβάνουν αριθμό από αναφορές και αντικείμενα , θέτοντας κάποια αναλογία μεταξύ αυτών κατά την δημιουργία.

Οι Papadakis et al. [PIN+12] χρησιμοποίησαν στην έρευνα τους δύο διαφορετικά σύνολα δεδομένων. Το σύνολο δεδομένων Dmovies, το οποίο περιλαμβάνει μια συλλογή ταινιών από τα σύνολα δεδομένων IMdb και DBpedia και το σύνολο Dinfoboxes ,το οποίο είναι το μεγαλύτερο σύνολο δεδομένων που αποτελείται από δύο διαφορετικές εκδοχές του DBpedia. Για την δημιουργία του έγιναν εξαγωγή όλα τα infoboxes των άρθρων της Wikipedia στην Αγγλική έκδοση. Τα infoboxes παρέχουν πληροφορίες για ένα άρθρο όπως χρονολογία δημοσίευσης, όνομα συγγραφέα, περιοδικό που δημοσιεύτηκε και πολλές άλλες πληροφορίες που σχετίζονται με ιδιότητες που χρησιμοποιούνται συνήθως. Όνομα ηθοποιού, όνομα ταινίας και εταιρεία παραγωγής χρησιμοποίησαν ως ιδιότητες και οι Homoceanu et al. [HKB14] στο δικό τους πείραμα. Οι πληροφορίες ανακτήθηκαν από σύνολα δεδομένων που προσφέρονται δωρεάν όπως DBpedia, και για κάθε σύνολο χρησιμοποίησαν διαφορετικό αριθμό ιδιοτήτων.

Χρησιμοποιώντας το σύστημα SLINT+ δημιούργησαν σύνολα που περιείχαν εγγραφές που αναφέρονταν στο ίδιο αντικείμενο με διαφορετικό τρόπο. Επέλεξαν μεν, κάθε φορά να γίνεται ταυτοποίηση τριών αντικειμένων. Οι Chen et al. [CJZ+12] παρουσίασαν την έρευνα τους τις πέντε πρώτες καλύτερες ιδιότητες από το σύνολο δεδομένων Cora, βάση της βαρύτητας που έχει η κάθε μία σε κάποια αξιολόγηση, με αύξων αριθμών κατάταξης. Και στις πέντε παρουσιάζεται η ιδιότητα συγγραφές, σε τρείς από αυτές η ιδιότητα τίτλος ως και επίσης παρουσιάζεται μέσα σε αυτές η ιδιότητα εκδότης, τόμος και περιοδικό. Χρησιμοποίησαν για το δικό τους πείραμα τις ιδιότητες συγγραφέας, τίτλος, διεύθυνση, ημερομηνία, σελίδα, εκδότης, τόμος, περιοδικό και συντάκτης. Οι Mirizzi et al. στο δικό τους πείραμα για την αξιολόγηση αλγορίθμων ταύτισης αντικειμένων χρησιμοποίησαν ιδιότητες που αναφέρονται σε ταινίες από την πηγή DBLP. Συγκεκριμένα χρησιμοποίησαν ιδιότητες όπως όνομα ηθοποιού, όνομα σκηνοθέτη, τίτλος ταινίας, χρονολογία παραγωγής και άλλες [MDR+12].

Στους Πίνακες 3.1 και 3.2 παρουσιάζονται οι ιδιότητες που έχουν χρησιμοποιηθεί καθώς και από ποιους συγγραφείς. Όπως παρατηρείται συνήθως χρησιμοποιούν σύνολα δεδομένων με ιδιότητες που έχουν να κάνουν με δημοσιεύσεις, δηλαδή, όνομα συγγραφέα, τίτλος άρθρου/δημοσίευσης, τοποθεσία και χρονολογία. Αυτές είναι οι ιδιότητες που εμφανίστηκαν κυρίως σε πειράματα. Επίσης η ιδιότητα movie και actor έχει χρησιμοποιηθεί αρκετές φορές από ερευνητές για πειράματα αξιολόγησης αλγορίθμων. Είναι οι συνήθεις ιδιότητες αλλά και οικογένειες διότι υπάρχουν πολλά σύνολα δεδομένων που μπορούν να χρησιμοποιηθούν ως πηγές. Επίσης, ο συγκεκριμένος τομέας, δηλαδή των δημοσιεύσεων και ταινιών, αυξάνονται συνεχώς λόγο των παραγωγών πολλών ταινιών αλλά και νέων δημοσιεύσεων από συγγραφείς.

Συγγραφείς Ιδιότητα	Yin et al.	Ioannou et al.	Mishra et al.	Kocphe et al	Dong et al.	Papadakis et al.	Chen et al.
Author	x	X	X	X	x	x	x
Title	x	X		X	x	x	x
Location	x	X	X		x		x
Publisher	x						x
Conference	x			X	x		x
Year	x	X	X	X	x	x	x
University		X					
Manufacturer				X			
Journal					x		x
Journal vol.					x		x

Πίνακας 3.1: Ιδιότητες που συνήθως χρησιμοποιούνται σχετικά με δημοσιεύσεις.

Συγγραφείς Ιδιότητα	Gu et al.	Herziq et al.	Papadakis et .al	Homoceanu et al.	Mirizzi et al.
Year		X	X		x
Name	X				
Age	X				
Journal					
Movie		X	X	x	x
Actor		X		x	x
manufacturer				x	x

Πίνακας 3.2: Ιδιότητες που συνήθως χρησιμοποιούνται σε σύνολα δεδομένων αναφορικά με ταινίες.

3.3 Καθορισμός Οικογενειών και Συνόλων

Σε αυτή την ενότητα θα γίνει καθορισμός των οικογενειών σύμφωνα με αναφορές που υπάρχουν στην ερευνητική κοινότητα. Οι πηγές (ακατέργαστων δεδομένων) από όπου θα ανακτηθούν οι καθορισμένοι πίνακες στήλες για κάθε οικογένεια είναι οι : IMDb, Amazon, DBLP, DbPedia. Οι συγκεκριμένες πηγές αποτελούνται από διάφορους

πίνακες με στήλες με διάφορες ιδιότητες όπως title, journal, name, book, university, movie κ.α. Σκοπός είναι να δημιουργηθούν οικογένειες από σύνολα δεδομένων που θα περιλαμβάνουν ιδιότητες που σχετίζονται μεταξύ τους και αναφέρονται στο ίδιο αντικείμενο, ώστε να είναι χρήσιμες στην ερευνητική κοινότητα για την αξιολόγηση αλγορίθμων ταύτισης αντικειμένων. Για κάθε οικογένεια που θα δημιουργηθεί θα χρησιμοποιηθεί κάποιος τροποποιητής, ο οποίος θα τροποποιεί τα δεδομένα ώστε να περιέχουν κάποιες παραλλαγές. Υπάρχουν πέντε κατηγορίες τροποποιητών που μπορούν να χρησιμοποιηθούν οι οποίοι είναι κατάλληλοι για ετερογενείς δεδομένα[IRV13].

Η πρώτη κατηγορία περιλαμβάνει συντακτικές παραλλαγές της πραγματικής τιμής ενός χαρακτηριστικού η του ονόματος του. Είναι συνέπεια των διαφορετικών τρόπων που μια αξία μπορεί να γραφτεί στην πραγματική ζωή χωρίς καμία αλλαγή της σημασίας της, ή το αποτέλεσμα ανθρώπινων λαθών. Περιλαμβάνει:

- Ορθογραφικά λάθη (misspellings)
- Ακρώνυμα, αρχικά και συντομογραφίες (acronyms, Initials and Abbreviations)
- Μετατροπές λέξεων (word permutations)
- Ψευδώνυμα και διαφορετικά πρότυπα (Aliases and Different Standards*)*
- Ομοιογένεια (Homonymity*)*

Η δεύτερη κατηγορία αναφέρεται στις δομικές παραλλαγές, οι οποίες μπορεί να υπάρχουν μεταξύ των χαρακτηριστικών του αντικειμένου. Αυτό έχει ως αποτέλεσμα δύο αντικείμενα που αντιπροσωπεύουν την ίδια εγγραφή στο πραγματικό κόσμο να διαφέρουν σημαντικά. Περιλαμβάνει:

- Χρήση πολλαπλών χαρακτηριστικών (Use of multiple attributes)
- Τιμές που λείπουν (Missing values)
- Μη καθορισμένα αντικείμενα (Underspecified entities)
- Εξειδικευμένα αντικείμενα (Overspecified entities)

Η τρίτη κατηγορία αναφέρεται σε σημαντικές παραλλαγές, που ακόμα και αν οι τιμές των ιδιοτήτων είναι ίδιες, το νόημα τους μπορεί να μην είναι. Αυτό συμβαίνει επειδή η ίδια λέξη χρησιμοποιείται συχνά για να εκπροσωπήσει διαφορετικά πράγματα. Το ίδιο ισχύει και την αντίθεση κατεύθυνση, δηλαδή διαφορετικές λέξεις μπορεί να αντιπροσωπεύουν την ίδια έννοια. Περιλαμβάνει:

- Συνώνυμα (Synonyms)
- Πολυγλωσσία (Multilingualism*)*

Η τέταρτη κατηγορία είναι η εξέλιξη των αντικειμένων που εξελίσσονται συνεχώς και δεν μένουν στατικά. Αυτή η εξέλιξη μπορεί να συνεπάγει αλλαγές στις τιμές των χαρακτηριστικών τους, την εξάλειψη χαρακτηριστικών, την προσθήκη νέων χαρακτηριστικών κλπ.

Τέλος στη πέμπτη κατηγορία , απόκλιση δικτύου σύνδεσης, τα αντικείμενα είναι συνδεδεμένα μεταξύ τους σχηματίζοντας ένα δίκτυο ενώσεων. Για να αναγνωρίζεται και να διακρίνεται ένα αντικείμενο από το άλλο, το δίκτυο τους παίζει σημαντικό ρόλο.

Για την δημιουργία των οικογενειών θα χρησιμοποιηθούν τροποποιητές που ανήκουν στην πρώτη κατηγορία, των συντακτικών παραλλαγών. Ο λόγος είναι ότι όλα τα σύνολα δεδομένων θα μπορούν να χρησιμοποιηθούν για την αξιολόγηση αλγορίθμων που ταυτοποιούν αντικείμενα σε σύνολα δεδομένων με συντακτικές παραλλαγές. Τα σύνολα όμως θα διαφέρουν σε ιδιότητες αλλά και σε παραμέτρους. Οι οικογένειες που θα δημιουργηθούν καθώς και ποιες ιδιότητες θα χρησιμοποιηθούν για την δημιουργία της κάθε μίας φαίνονται στο πίνακα 3.1. Η δημιουργία των οικογενειών έγινε λαμβάνοντας υπόψη τους συνδυασμούς που χρησιμοποιούνται συνήθως στην ερευνητική κοινότητα για αξιολόγηση αλγορίθμων ταύτισης αντικειμένων. Για να γίνει σωστή ταυτοποίηση των αντικειμένων οι ιδιότητες πρέπει να σχετίζονται μεταξύ τους και να αναφέρονται στην ίδιο θέμα, π.χ. ένα άρθρο γράφτηκε από ένα συγγραφέα, το όνομα συγγραφέα και κάποιο άρθρο συσχετίζονται. Θα ήταν λάθος να δημιουργηθούν οικογένειες συνόλων δεδομένων που οι ιδιότητες τους δεν σχετίζονται γιατί θα υπήρχε αναφορά σε διαφορετικά θέματα. Δηλαδή δεν μπορεί να χρησιμοποιηθεί η ιδιότητα συγγραφέας με την ιδιότητα εστιατόριο στην ίδια οικογένεια. Για αυτό και έγινε αναφορά σε υφιστάμενες μελέτες για να υπάρχει σωστό δείγμα οικογενειών, ώστε οι οικογένειες που θα δημιουργηθούν να έχουν ομοιότητες με αυτές. Για την πρώτη οικογένεια χρησιμοποιήθηκαν ιδιότητες που αναφέρονται στο θέμα των ταινιών. Χρησιμοποιήθηκαν οι ιδιότητες name,movie,film studio,manufacturer και year και έγινε δημιουργία των Collection A1,A2,A3,A4,A5 και A6. Η δεύτερη συλλογή δεδομένων αναφέρεται σε ερευνητές και περιέχει τις ιδιότητες name,university,research,born year και death(αν ο ερευνητής απεβίωσε θα εμφανίζεται η τιμή Null, αν ζει θα εμφανίζεται η τιμή 1. Για την δεύτερη οικογένεια δημιουργήθηκα τα Collection B1, B2 και B3. Η τρίτη οικογένεια αναφέρεται σε δημοσιεύσεις και περιλαμβάνει τις ιδιότητες author, book

title, publisher, publication paper, publication journal , publication series και year. Για την τρίτη οικογένεια δημιουργήθηκαν τα σύνολα δεδομένων C1, C2, C3, C4 και C6. Για την οικογένεια D και E και τα σύνολα δεδομένων αυτών, χρησιμοποιήθηκαν οι ίδιες ιδιότητες που χρησιμοποιήθηκαν στην οικογένεια B.

Επίσης, η κάθε οικογένεια και τα σύνολα δεδομένων αυτής δημιουργηθήκαν ώστε να αξιολογούν ένα αλγόριθμο ταύτισης αντικειμένων ως προς την ποιότητα, το χρόνο που χρειάζεται να ταυτοποιήσει και την συμπεριφορά αυτού όσο αυξάνετε ο αριθμός των συνόλων δεδομένων(scalability). Οπόταν, κατά την δημιουργία των συνόλων δεδομένων όλες οι τροποποιήσεις και οι παράμετροι διαμορφώθηκαν και εφαρμόστηκαν έτσι ώστε να υπάρχει ένα αξιοπρεπές δείγμα που θα μπορεί να χρησιμοποιηθεί για την αξιολόγηση αλγορίθμων ταύτισης αντικειμένων. Η πρώτη οικογένεια θα αξιολογεί την ποιότητα ενός αλγορίθμου και την συμπεριφορά αυτού καθώς ο αριθμός των αντικειμένων αυξάνετε. Για αυτό για την πρώτη οικογένεια η μόνη παράμετρος που διαφοροποιείται ανάμεσα στα Collections είναι ο αριθμός των παρατηρήσεων όπως φαίνεται και στο πίνακα 4.3. Σκοπός είναι να υπάρχουν Collections με τις ίδιες παραμέτρους αλλά κάθε φορά με αυξημένο αριθμό παρατηρήσεων έτσι ώστε να μπορεί να αξιολογηθεί η ικανότητα του αλγορίθμου να ταυτοποιεί σε σύνολα δεδομένων με μεγαλύτερο αριθμών αντικειμένων. Επίσης θα αξιολογεί τον χρόνο που χρειάζεται για την ταυτοποίηση ένας αλγόριθμός όσο αυξάνεται ο αριθμός των αντικειμένων.

Για τα Collections της δεύτερης οικογένειας χρησιμοποιήθηκαν αρκετοί διαφορετικοί συνδυασμοί παραμέτρων και ιδιοτήτων έτσι ώστε να γίνεται αξιολόγηση για την ικανότητα ενός αλγορίθμου να ταυτοποιεί αντικείμενα σε διαφορετικές καταστάσεις. Ο αριθμός των παρατηρήσεων αυτή την φορά επιλέχθηκε να παραμένει ο ίδιος έτσι ώστε να υπάρχουν σύνολα δεδομένων που έχουν τον ίδιο αριθμών παρατηρήσεων αλλά διαφέρουν ως προς τα ιδιότητες και τις διάφορες παραμέτρους που μπορούν να δοθούν όπως φαίνεται και στον πίνακα 4.3. Επίσης θα μπορούν να αξιολογηθούν αλγόριθμοι ταύτισης αντικειμένων ως προς τον χρόνο που χρειάζονται για την ταυτοποίηση σε σύνολα δεδομένων διαφορετικών καταστάσεων κάθε φορά.

Η τρίτη οικογένεια επιλέχθηκε να είναι η πιο πολύπλοκη. Εδώ αυξάνεται και ο αριθμός των παρατηρήσεων αλλά και διαφοροποιούνται και οι ιδιότητες αλλά και οι παράμετροι. Οπόταν αλγόριθμοι ταύτισης αντικειμένων θα μπορούν να αξιολογηθούν ως προς την ικανότητα ταυτοποίησης και την συμπεριφορά τους καθώς αυξάνεται ο

αριθμός αντικειμένων αλλά παράλληλα υπάρχουν διαφορές στις παραμέτρους των συνόλων δεδομένων. Επίσης θα αξιολογείτε αλλά και ο χρόνος που χρειάζονται να ταυτοποιήσουν αντικείμενα σε σύνολα δεδομένων που διαφέρουν εξ ολοκληρωτικά μεταξύ τους. Τι διαφορές έχουν τα Collections μεταξύ τους παρουσιάζεται στο πίνακα 4.3.

Τα Collection D και E δημιουργήθηκαν με σκοπό να χρησιμοποιηθούν σε περίπτωση που κατά την αξιολόγηση ενός αλγορίθμου στα Collections B και C, παρουσιαστεί κάποιο πρόβλημα. Στα Collections B και C επειδή τα σύνολα δεδομένων που τα αποτελούν διαφέρουν αρκετά ως και όλες τις παραμέτρους, δεν θα μπορεί να κατατοπιστεί που ακριβώς είναι το πρόβλημα που παρουσιάστηκε κατά την αξιολόγηση ενός αλγορίθμου. Για αυτό στο Collection D το μόνο που διαφοροποιείται ανάμεσα στα σύνολα δεδομένων του είναι η αναλογία των διπλότυπων έναντι του αριθμού των παρατηρήσεων. Οπόταν αν χρησιμοποιηθούν τα Collections D1, D2 και D3 και παρουσιαστεί πρόβλημα πάλι κατά την αξιολόγηση ενός αλγορίθμου κατά την ταυτοποίηση τότε το πρόβλημα σημαίνει θα είναι σχετικό με την αναλογία των διπλότυπων έναντι του αριθμού των παρατηρήσεων. Με το ίδιο σκεπτικό δημιουργήθηκε και το Collection E, και τα σύνολα δεδομένων αυτού . Τα Collections E1,E2 και E3 διαφέρουν μεταξύ τους μόνο στο ποσοστό του τροποποιητή που χρησιμοποιήθηκε. Οπόταν αν ένας αλγόριθμος παρουσιάσει πρόβλημα κατά την διαδικασία ταύτισης αντικειμένων , μπορεί να χρησιμοποιηθούν ένα από τα Collections E1,E2 και E3, έτσι ώστε αν παρουσιαστεί πρόβλημα ξανά και την διαδικασία, αφού στο μόνο που διαφέρουν είναι το ποσοστό του τροποποιητή, λογικό θα είναι το πρόβλημα να παρουσιάζεται στην συγκεκριμένη παράμετρο.

Οικογένειες	Σύνολα Δεδομένων	Ιδιότητες
A	A1,A2,A3,A4,A5,A6	Actor/Name
		Movie
		Film studio
		Manufacturer
		Year
B	B1,B2,B3	Researcher/Name
		University
		Research
		Born Year
		Age of Death
		Alive(Yes or No)
C	C1,C2,C3,C4	Author/Name
		Book title
		Publisher
		Publication Paper
		Publication Journal
		Publication Series
		Year
D	D1,D2,D3	Researcher/Name
		University
		Research
		Born Year
		Age of Death
		Alive(Yes or No)
E	E1,E2,E3	Researcher/Name
		University
		Research
		Born Year
		Age of Death
		Alive(Yes or No)

Πίνακας 3.1: Καθορισμός οικογενειών με τις αντίστοιχες ιδιότητες τις κάθε μιας.

3.4 Συνεισφορά της εν λόγω εργασίας

Τα υφιστάμενα σύνολα δεδομένων όπως αναφέρθηκε και στην εισαγωγή μπορεί να έχουν δημιουργηθεί μέσω της συνάθροισης περισσότερων από ένα είδος ετερογένειας, με αποτέλεσμα να μην είναι σαφές ακόμη και σε ένα έμπειρο ερευνητή τι είδους είναι αυτές οι ετερογένειες.

Τα υφιστάμενα σύνολα δεδομένων που έχουν δημιουργηθεί από ερευνητές, βασίζονται σε δοκιμαστικά έτοιμα σύνολα που δεν είναι σαφής η ετερογένεια τους. Ακόμη και αν δημιούργησαν καινούργια αυτά υστερούν είτε στον αριθμό ιδιοτήτων, είτε δεν είχαν κάποια συντακτική παραλλαγή είτε ο αριθμός των παρατηρήσεων είναι πολύ μικρός για να υπάρξει σωστή αξιολόγηση.

Στην εργασία αυτή θα δημιουργηθούν σύνολα τα οποία περιέχουν περισσότερες από μια ιδιότητες καθώς επίσης και επαρκή αριθμό δείγματος ή παρατήρησης έτσι ώστε να μπορούν να αξιολογηθούν και να ταυτοποιηθούν στην επανάληψη. Επιπλέον ένα πολύ σημαντικό στοιχείο που δεν έχει σχεδόν συμπεριληφθεί στις αναφερόμενες μελέτες είναι οι συντακτικές παραλλαγές. Η εργασία αυτή έχει ως στόχο να συμπεριλάβει και αυτή τη παράμετρο στα σύνολα που θα δημιουργηθούν.

Κεφάλαιο 4

Μεθοδολογία της Δημιουργίας των Συνόλων με Δεδομένα

Κύριος σκοπός αυτής της μεταπτυχιακής διατριβής είναι να γίνει μια λεπτομερής και εμπεριστατωμένη παρουσίαση για τον καθορισμό των οικογενειών, μέσα από υφιστάμενη βιβλιογραφία αλλά και χρήσης πειραμάτων για την δημιουργία των οικογενειών. Επιπλέον, σκοπός της ήταν η δημιουργία συνόλων από δεδομένα, χρησιμοποιώντας το σύστημα EMBench, τα οποία θα μπορούν να χρησιμοποιηθούν για την αξιολόγηση αλγορίθμων ταύτισης αντικειμένων. Αρχικά θα γίνει παρουσίαση του συστήματος και των λειτουργιών που έχει και στην συνέχεια μέσα από το πείραμα θα γίνει δημιουργία των οικογενειών που καθοριστήκαν χρησιμοποιώντας το σύστημα.

Για την δημιουργία των οικογενειών εξετάστηκαν θέματα όπως η ταυτοποίηση αντικειμένων, τα συστήματα ταυτοποίησης που χρησιμοποιούνται συνήθως, όπως τα ALIAS, PIM, SOCCER, ENS, SWING, ως επίσης και οι ιδιότητες που χρησιμοποιούν οι ερευνητές για την δημιουργία συνόλων δεδομένων για χρήση ταύτισης αντικειμένων. Το σύστημα που θα χρησιμοποιηθεί είναι το EMBench που όπως αναφέρθηκε στο κεφάλαιο δύο προσφέρει την δυνατότητα στον χρήση να διαμορφώσει όπως αυτός θέλει τα σύνολα δεδομένα που θα δημιουργηθούν.

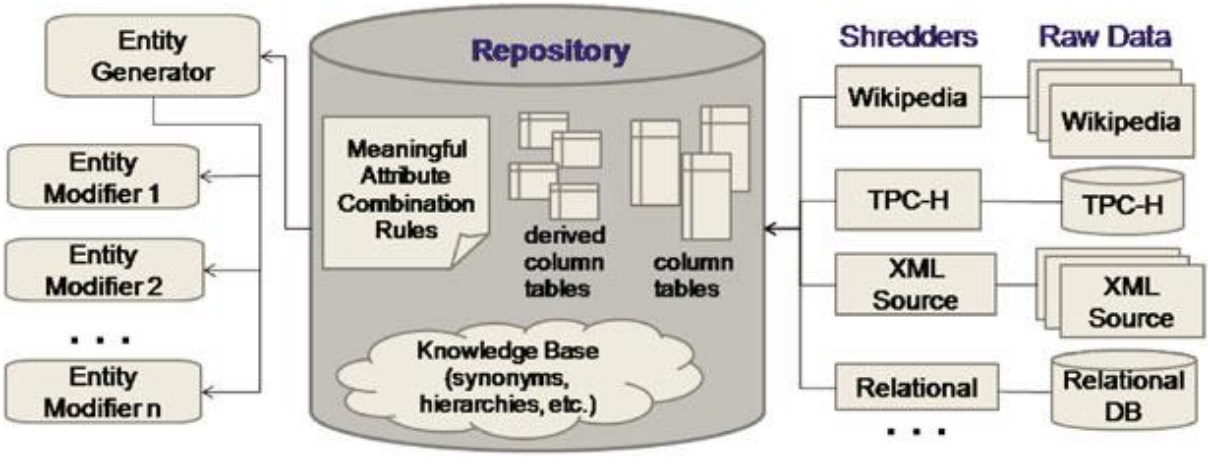
4.1 Το Σύστημα EMBench

Όπως έχει αναφερθεί, το EMBench είναι σύστημα αξιολόγησης αλγορίθμων ταύτισης αντικειμένων[IRV13]. Κύριος σκοπός του είναι η παραγωγή συνθετικών δεδομένων για την αξιολόγηση αλγορίθμων ταυτοποίησης αντικειμένων, ωσότου μπορεί να χρησιμοποιηθεί για δύο επιπλέον σκοπούς. Ο ένας είναι η δημιουργία ενός χώρου δεδομένων από πραγματικές πηγές δεδομένων που μπορεί να γίνει με τη χρήση των τροποποιητών. Ο δεύτερος είναι η δημιουργία μεγάλων συνθετικών συλλογών

ταυτοποίησης, μαζί με διάφορους τύπους αντικειμένων, για τα οποία δεν υπάρχουν τροποποιητές για να εφαρμοστούν.

Με το σύστημα EMBench ο χρήστης μπορεί να διαμορφώσεις τις παραμέτρους όπως αυτός θέλει με αποτέλεσμα την δημιουργία πολύπλοκων συνόλων δεδομένων για την αξιολόγηση ενός αλγορίθμου ταυτοποίησης αντικειμένων. Η γενική αρχιτεκτονική του συστήματος απεικονίζεται στο σχήμα 4.1. Αποτελείται από τρία κύρια μέρη: (i) από ένα χώρο από δεδομένα που θα χρησιμοποιηθούν στην κατασκευή των συλλογών αντικειμένων όπως αυτό επιλέχθηκε από τον χρήστη , παράλληλα με τα στοιχεία που ονομάζονται τεμαχιστές που διαμορφώνουν αυτά τα δεδομένα ώστε να είναι κατάλληλα και ποιοτικά. (ii) Από μια μηχανή παραγωγής ταυτοποίησης που συνθέτει τα αντικείμενα στους χώρους δεδομένων για να διατυπώσει μια συλλογή από ταυτοποιήσεις και (iii) από μια σειρά από τροποποιητές που τροποποιούν με διάφορους τρόπους τα δεδομένα για χρήση αξιολόγησης ταυτοποίησης και κατασκευάζουν μια νέα συλλογή ετερογενών αντικειμένων.

Το EMBench προσφέρει τρεις βασικές λειτουργίες[IRV13]. Η πρώτη είναι η δημιουργία ενός χώρου που θα είναι η πηγή που θα περιέχει τις παρατηρήσεις, εισάγοντας δεδομένα χρησιμοποιώντας τους τεμαχιστές. Η δεύτερη λειτουργία είναι η δημιουργία συνόλων αντικειμένων χρησιμοποιώντας τα δεδομένα από τις πηγές των παρατηρήσεων και η τρίτη λειτουργία είναι η αξιολόγηση αλγορίθμων ταύτισης αντικειμένων.



Σχήμα 4.1: Η γενική αρχιτεκτονική του EMBench παράλληλα με τη ροή των δεδομένων μεταξύ των διαφόρων συνιστωσών[IRV13].

4.2 Λειτουργία του Συστήματος EMBench

Στο πρώτο μέρος του συστήματος, όπου είναι ο χώρος δεδομένων, γίνεται η συλλογή των δεδομένων. Εδώ, ο τεμαχιστής παίρνει ένα σύνολο παρατηρήσεων και το διαχωρίζει σε μια σειρά από πίνακες με στήλες. Ο χρήστης έχει την δυνατότητα να επιλέξει ποια σύνολα δεδομένων θα διαχωριστούν, και προσθέτει επιπρόσθετα σύνολα χρειάζεται χρησιμοποιώντας τον εξειδικευμένο τεμαχιστή ή χρησιμοποιώντας ένα γενικό που υπάρχει ήδη στο σύστημα. Υπάρχει όμως η πιθανότητα δύο διαφορετικές πηγές να περιέχουν πληροφορίες με τα ίδια δεδομένα, π.χ. ονόματα πόλεων[IRV13]. Μέσα από διαφορετικούς τεμαχιστές θα δημιουργηθούν δύο διαφορετικοί πίνακες από στήλες. Σε περίπτωση όμως που επιθυμείται να υπάρχει μόνο ένα τομέας με ονόματα πόλεων, τότε οι δύο πίνακες θα συγχωνευτούν σε ένα. Ως αποτέλεσμα θα είναι η εξάλειψη των πολλαπλών τιμών που υπάρχουν μεταξύ τους. Είναι πιθανό οι τιμές ενός πίνακα με στήλες να χωριστούν σε διαφορετικού πίνακες, παραδείγματος χάρη σε περίπτωση ονόματος και επιθέτου θα δημιουργηθεί ένας πίνακας με τα ονόματα και ένας με τα επίθετα. Για την αντιμετώπιση τέτοιου είδους προβλημάτων στο EMBench συμπεριλαμβάνεται ένα σύνολο προκαθορισμένων κανόνων που μπορούν να επεκταθούν περαιτέρω από τον χρήστη, και οι οποίοι κανόνες αυτοί καθορίζουν πως οι τιμές των στηλών του πίνακα πρόκειται να συνδυαστούν μεταξύ τους ή τροποποιημένα και να κατευθύνουν τη δημιουργία ενός νέου δείγματος στηλών, που θα ονομάζονται ως παράγωγες στήλες του πίνακα.

Στο δεύτερο μέρος του συστήματος (Μηχανή παραγωγής Ταυτοποίησης) αφού έχουν δημιουργηθεί οι τομείς και οι παράγωγοι πίνακες, μπορεί να παραχθεί η αρχική συλλογή αντικειμένων. Η συλλογή αντικειμένων παράγεται με τη δημιουργία δεδομένω με χαρακτηριστικά που επιλέγονται από τους παράγωγους πίνακες (το όνομα του χαρακτηριστικού είναι το όνομα του παράγωγου πίνακα και η τιμή είναι μια από τις τιμές του). Ωστόσο, δεν πρέπει να δημιουργηθεί εντελώς τυχαία η συλλογή αντικειμένων, αλλά να υπάρχει κάποιος έλεγχος κατά την δημιουργία. Για το λόγο αυτό αποφασίζεται ο αριθμός των αντικειμένων που χρειάζεται να δημιουργηθούν και ο μέγιστος αριθμών χαρακτηριστικών που αναμένεται να έχουν τα αντικείμενα. Ακολούθως πραγματοποιείτε τυχαία επιλογή από ιδιότητες από τον χώρο των παράγωγων πινάκων, και κάθε φορά επιλέγεται ένας παράγωγος πίνακας με στήλες. Για να μην επιτραπεί στα αντικείμενα να έχουν πολλαπλές τιμές από χαρακτηριστικά, πρέπει να διασφαλιστεί ότι οι ιδιότητες που επιλεχθήκαν δεν επαναλαμβάνονται, π.χ.

ποτέ δεν γίνεται επιλογή του ίδιου παράγωγου πίνακα στηλών δύο φορές. Μετέπειτα δημιουργείται ένα σχεσιακός πίνακας ο οποίος συμπληρώνεται με δεδομένα επιλέγοντας για κάθε αντικείμενο, τις τιμές που χρειάζεται από τον αντίστοιχο παράγωγο πίνακα. Υπάρχουν δύο επιλογές κατά την διαδικασία των τιμών που χρειάζεται από τον πίνακα στηλών. Είτε να γίνεται πάντα μια τυχαία επιλογή, πράγμα που σημαίνει ότι μπορεί να υπάρξουν κάποιες επαναλήψεις, ή να γίνεται μια τυχαία επιλογή τιμών σύμφωνα με την διανομή Zipfian.

Στο τρίτο μέρος (Τροποποιητές Αντικειμένων) γίνεται δημιουργία του τροποποιημένου συνόλου δεδομένων. Αυτό επιτυγχάνεται περνώντας τον πίνακα μέσα από μια σειρά τροποιητών. Η είσοδος σε ένα τροποποιητή είναι ένας πίνακας και η έξοδος είναι ένας πίνακας με το ίδιο σχήμα και αριθμών πλειάδων αλλά με τροποποιημένες τιμές. Στο τέλος της διαδικασίας, ο τροποποιημένος πίνακας , χρησιμοποιείται για την παραγωγή μια συλλογή ταυτοποίησης που παίζει το ρόλο της τροποποιημένης συλλογής ταυτοποίησης .

Με τις αρχικές και τροποποιημένες συλλογές αντικειμένων έτοιμες μπορουν για δημιουργηθούν τα σύνολα δεδομένων. Για την εισαγωγή συνόλων δεδομένων υπάρχει η επιλογή είτε να χρησιμοποιηθούν δεδομένα που υπάρχουν στην προεπιλεγμένη εφαρμογή του EMBench όπως φαίνονται στο σχήμα 4.1 ή να συλλεχτούν δεδομένα από τις υπάρχουσες πηγές το οποίο αυτό μπορεί να επιτευχθεί με τη χρήση των τεμαχιστών. Επιλέγοντας ποιοι τροποποιητές θα ενεργοποιηθούν γίνεται έλεγχος ποιες πηγές πρέπει να αντιγραφούν και ποιες όχι. Για να πραγματοποιηθεί η συλλογή δεδομένων και η τροποποίηση πρέπει απλά να επιλεχτεί ποιοι τροποποιητές πρέπει να χρησιμοποιηθούν, ως και επίσης ποιές παράμετροι θα διαμορφωθούν στις πηγές από όπου θα ανακτηθούν τα δεδομένα.

Όταν τα συλλεγμένα δεδομένα είναι στη θέση τους, πρέπει να καθοριστεί ο τρόπος που θα συνδυαστούν οι πίνακες στηλών για να δημιουργηθεί η αρχική και η τροποποιημένη συλλογή αντικειμένων.. Το αποτέλεσμα θα είναι ένα σύνολο συλλογών αντικειμένων με κάθε συλλογή να συνοδεύεται με τη τροποποιημένη εκδοχή. Έτσι, κάθε αντικείμενο έχει δύο εκδοχές: η μία που περιέχει τα αρχικά δεδομένα και μια άλλη που περιέχει τα τροποποιημένα δεδομένα. Επιπλέον θα δημιουργηθεί μια λίστα που παρέχει μια ακολουθία από τροποποιητές οι οποίοι εφαρμόστηκαν σε κάθε αντικείμενο.

4.3 Δημιουργία των Συνόλων με Δεδομένα

Για την δημιουργία των οικογενειών χρησιμοποιήθηκε laptop σε περιβάλλον Windows 10 64-bit με επεξεργαστή Intel Core i3-4000M στα 2.40GHZ και μνήμη RAM 6.00GB. Επίσης χρησιμοποιήθηκε η τελευταία εκδοχή Java και η MySQL για την δημιουργία των πινάκων. Χρειάστηκαν 4104.573 δευτερόλεπτα για την δημιουργία των συλλογών, και 8206.816 δευτερόλεπτα για να εκτελεστούν οι τροποποιητές.

Χρησιμοποιήθηκε ένα configuration file που περιλάμβανε όλες τις παραμέτρους και έτρεξε μια φορά μόνο. Για κάθε πίνακα ξεχωριστά, δόθηκαν οδηγίες ποιες ιδιότητες θα χρησιμοποιηθούν, πως θα ονομάζονται οι στήλες του κάθε πίνακα, ποιες ιδιότητες από τα σύνολα παρατηρήσεων θα υπάρχουν σε κάθε πίνακα, ποιος τροποποιητής θα χρησιμοποιηθεί και σε τι ποσοστό και ποια θα είναι η αναλογία των διπλότυπων έναντι του συνόλου των παρατηρήσεων. Οι οικογένειες που θα δημιουργηθούν καθώς και οι ιδιότητες για κάθε μια από αυτές παρουσιάστηκαν ήδη στο πίνακα 3.1 του κεφαλαίου 3. Οι παράμετροι που χρησιμοποιηθήκαν καθώς και τα ποσοστά παρουσιάζονται στο πίνακα 4.3. Κατά την δημιουργία των κανόνων και των ιδιοτήτων στον τεμαχιστή, λήφθηκαν υπόψη οι κανόνες των ιδιοτήτων που προϋπήρχαν σε κάθε σύνολο δεδομένων. Βάση αυτών έγιναν οι κατάλληλες ρυθμίσεις στο τεμαχιστή ώστε να δημιουργηθούν οι σωστές οικογένειες με τα σωστά σύνολα και τις σωστές τιμές. Στον πίνακα 4.2 που ακολουθεί παρουσιάζονται αυτοί οι κανόνες για κάθε σύνολο παρατηρήσεων ξεχωριστά.

Βάση Δεδομένων(Πηγή)	Κανόνες Δεδομένων(dataRules)
Amazon	Company = $FilmManufacturer;
	Company = $FilmStudio;
	Occupation = $EntertainmentRelatedOccupation
BabyNames	BabyName = $FeminineBabyName;
	BabyName = $MasculinBabyName
DBLP	Company = $ScientificPublisher;
	Book = $ComputerScienceBookTitle
DbPedia	Landmark = $University; Landmark = $Theatre; Landmark = $Museum; Landmark = $Mountain; Landmark = $Monastery; Company = $Newspaper
IMDb	Company = $FilmStudio; Occupation = $EntertainmentRelatedOccupation; FirstName = $FeminineFirstName; FirstName = $MasculinFirstName; Name = $FirstName + ' ' + $Surname; NameWithTitle = 'Mr. ' + $MasculinFirstName + ' ' + $Surname; NameWithTitle = 'Ms. ' + $FeminineFirstName + ' ' + $Surname; Company = $FilmManufacturer

Πίνακας 4.2: Παρουσίαση των κανόνων για την κάθε βάση δεδομένων.

Οπόταν τα Collections όπως έχει ήδη αναφερθεί , επιλέχθηκαν να διαφέρουν είτε ως προς αριθμό παρατηρήσεων, των αριθμών ιδιοτήτων, τον τροποποιητή κτλ. Τα Collections A1, A2, A3, A4, A5 και A6 στο μόνο που θα διαφέρουν είναι ο αριθμός των παρατηρήσεων. Κάθε φορά για κάθε νέο Collection που θα δημιουργηθεί θα αυξάνεται και αντίστοιχα ο αριθμός των παρατηρήσεων. Για το Collection A1 θα υπάρχουν 10000 παρατηρήσεις, για το A2 15000, για το A3 20000, για το A4 25000, για το A5 30000 και τέλος για το A6 35000. Το ότι αυξάνεται κάθε φορά ο αριθμός των παρατηρήσεων ανά πέντε χιλιάδες είναι κάτι εντελώς τυχαίο και δεν αποσκοπεί σε κάποια συγκεκριμένη παρατήρηση.

Απλά θα ήταν καλό να υπάρχει μια σχετική καλή διαφορά μεταξύ των αριθμών των παρατηρήσεων των Collections έτσι ώστε όταν θα γίνει αξιολόγηση αλγορίθμων ταύτισης αντικειμένων με αυτά τα Collections να υπάρχει μια σεβαστή διαφορά παρατηρήσεων για να αξιολογηθεί ο αλγόριθμος για το χρόνο που χρειάστηκε για την διαδικασία ταύτισης. Για τα Collections της οικογένειας Α, επιλέχθηκε να έχουν όλα αριθμών ιδιοτήτων 11. θα χρησιμοποιηθεί ο τροποποιητής για ορθογραφικά λάθη με ποσοστό 6% και η αναλογία των διπλότυπων έναντι του συνόλου των παρατηρήσεων θα είναι της τάξης του 9%.

Για τα Collections B1, B2 και B3 θα χρησιμοποιηθεί ο ίδιο αριθμός παρατηρήσεων μιας και θα χρησιμοποιούνται για την αξιολόγηση της συμπεριφοράς αλγορίθμων σσε ίδιο αριθμών παρατηρήσεων αλλά με διαφορετικές παραμέτρους και ιδιότητες κάθε φορά. Οπόταν ο αριθμός των παρατηρήσεων θα παραμένει σταθερός κάθε φορά στις 34000. Ο αριθμός των ιδιοτήτων που θα χρησιμοποιούνται κάθε φορά θα είναι διαφορετικός. Οι τροποποιητές που θα χρησιμοποιηθούν θα είναι τα ορθογραφικά λάθη και οι μετατροπές λέξεων με ποσοστό 2% , 5% και 10% έκαστος για το Collection B1, B2 και B3 αντίστοιχα. Επίσης διαφορετική θα είναι κάθε φορά η αναλογία των διπλότυπων έναντι του συνόλου των παρατηρήσεων, για το Collection B1 6%, για το B2 8% και τέλος για το B3 10%.

Τέλος για την δημιουργία της οικογένειας C, θα χρησιμοποιηθούν διάφοροι συνδυασμοί για όλες τις παραμέτρους κάθε φορά. Η οικογένεια αυτή είναι και η πιο πολύπλοκη μιας και θα αυξάνεται ο αριθμός των παρατηρήσεων για κάθε Collection που θα δημιουργείται ως και επίσης θα διαφοροποιείται μια παράμετρος. Αναλυτικά ο αριθμός των παρατηρήσεων για τα Collections C1, C2, C3, C4, C5 και C6 θα είναι 8000,12000,16000,20000,24000 και 28000 αντίστοιχα. Για τα C1, C2, C3 και C4 θα χρησιμοποιηθούν για τις τροποποιήσεις οι τροποποιητές για ακρώνυμα και μετατροπών λέξεων με ποσοστά 4% , 6% , 8% και 10% αντίστοιχα. Για τα C5 και C6 θα χρησιμοποιηθεί ο τροποποιητής για ορθογραφικά λάθη με ποσοστό 4% και 8% αντίστοιχα. Η αναλογία διπλοτύπων έναντι του συνολικού αριθμού των παρατηρήσεων θα είναι για το C1 4%, για C2 το 6%, για C3 το 8% ,για το C4 6% , για το C5 4% και για το C6 8%. Τέλος ο αριθμός των ιδιοτήτων που θα χρησιμοποιηθεί θα είναι επτά για το Collection C1, στην συνέχεια θα αυξηθεί σε οχτώ για το Collection C2, μετά θα αυξηθεί εκ νέου σε εννέα για το Collection C3 και θα αυξηθεί ξανά δέκα για τα Collections C4,C5 και C6.

Τα Collections D1 ,D2, D3 και E1, E2, E3 έχουν τις ίδιες ιδιότητες με τα Collections της οικογένειας B. Για τα D1,D2 και D3 τροποποιείται η παράμετρος αναλογία διπλοτύπων έναντι του συνολικού αριθμού παρατηρήσεων ενώ οι άλλοι παράμετροι παραμένουν σταθερές. Στα Collections E1, E2 και E3 τροποποιείται μόνο το ποσοστό του τροποποιητή που χρησιμοποιήθηκε. Οι υπόλοιποι παράμετροι παραμένουν σταθερές. Όπως αναφέρθηκε και στο κεφάλαιο 3, ο λόγος δημιουργία των Collection D1, D2, D3, E1, E2 και E3 είναι να υπάρχουν σύνολα δεδομένα με διαφοροποιημένη μόνο μια παράμετρο, για εντοπισμό προβλήματος κατά την αξιολόγηση αλγορίθμων ταύτισης αντικειμένων. Για παράδειγμα αν ένας αλγόριθμος παρουσιάσει πρόβλημα κατά την ταυτοποίησης στα Collections της οικογένειας C, μπορούν να χρησιμοποιηθούν τα Collections των οικογενειών D και E, για τον εντοπισμό αυτού του προβλήματος.

Συλλογή	Αριθμός Ιδιοτήτων	Αριθμός Παρατηρήσεων	Τροποποιη-τής/τές	Ποσοστό Τροποποιη-τή/τών	Ratio of Duplica-tion vs Total
A1	11	10000	Misspelling	6%	9%
A2	11	15000	Misspelling	6%	9%
A3	11	20000	Misspelling	6%	9%
A4	11	25000	Misspelling	6%	9%
A5	11	30000	Misspelling	6%	9%
A6	11	35000	Misspelling	6%	9%
B1	8	34000	Misspelling/ Permutation	2%/2%	6%
B2	9	34000	Misspelling/ Permutation	5%/5%	8%
B3	10	34000	Misspelling/ Permutation	10%/10%	10%
C1	7	8000	Acronym/ Permutation	4%/4%	4%
C2	8	12000	Acronym/ Permutation	6%/6%	6%
C3	9	16000	Acronym/	8%/8%	8%

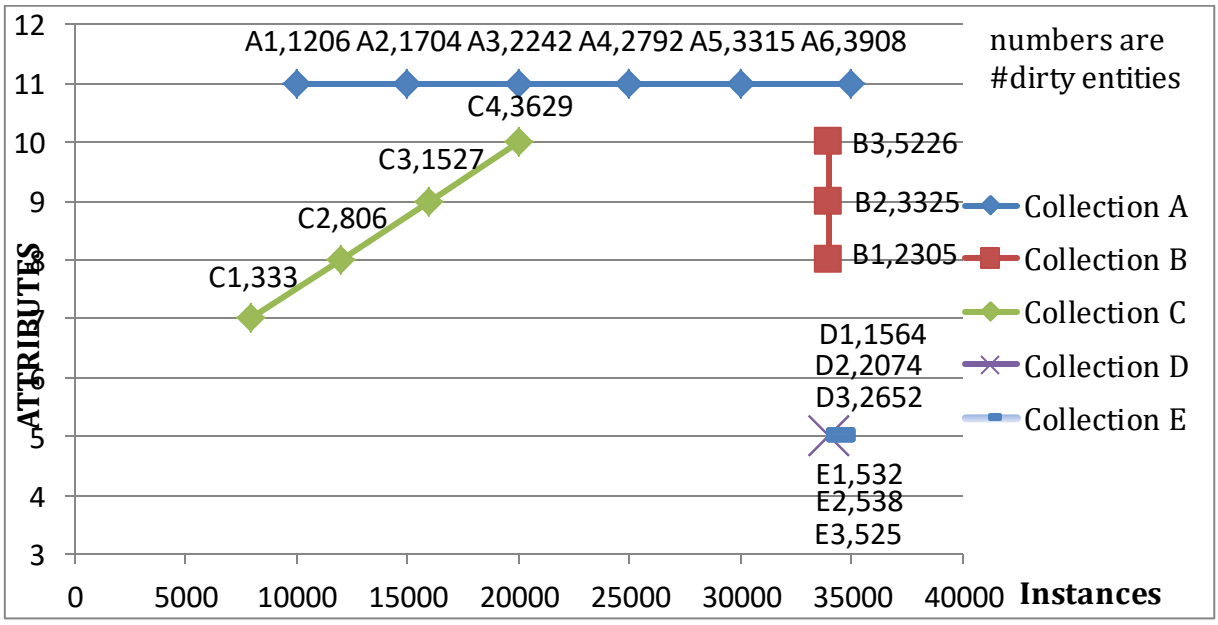
			Permutation		
C4	10	20000	Acronym/ Permutation	10%/10%	10%
D1	5	34000	Misspelling/ Permutation	2%/2%	6%
D2	5	34000	Misspelling/ Permutation	2%/2%	8%
D3	5	34000	Misspelling/ Permutation	2%/2%	10%
E1	5	34000	Misspelling/ Permutation	2%/2%	2%
E2	5	34000	Misspelling/ Permutation	5%/5%	2%
E3	5	34000	Misspelling/ Permutation	10%/10%	2%

Πίνακας 4.3.: Παρουσίαση συνόλων δεδομένων που θα δημιουργηθούν με τις παραμέτρους για κάθε ένα.

Κεφάλαιο 5

Ανάλυση Δεδομένων, Αποτελέσματα και Συζήτηση

Οι οικογένειες που δημιουργήθηκαν όπως έχε ήδη αναφερθεί διαφέρουν η κάθε μια μεταξύ τους σε κάποιες παραμέτρους ώστε να υπάρχουν διαφορετικές συλλογές για την αξιόλογη αλγορίθμων ταύτισης αντικειμένων. Επίσης υπάρχουν πολύπλοκα σύνολα δεδομένων, υπάρχουν και κάποια που είναι πιο «απλά», και κάποια άλλα που αλλάζει ή μένει σταθερή κάποια παράμετρος ώστε να μπορούν να χρησιμοποιηθούν σε διάφορων ειδών πειραμάτων αξιολόγησης αλγορίθμων ταύτισης αντικειμένων αναλόγως με την περίπτωση. Στο διάγραμμα 5.1 παρουσιάζονται τα αποτελέσματα της δημιουργία των οικογενειών ανά Collections. Στον άξονα instances είναι ο αριθμός των παρατηρήσεων που χρησιμοποιήθηκε για κάθε Collection και στον άξονα attributes είναι ο αριθμός ιδιοτήτων που χρησιμοποιήθηκε για κάθε Collection. Οι αριθμοί που εμφανίζονται σε κάθε σημείο είναι ο αριθμός των παρατηρήσεων που τροποποιήθηκαν, «dirty» .



Διάγραμμα 5.1: Αποτελέσματα Collection's A, B, C,D,E

Όπως παρατηρείται στο διάγραμμα 5.1 για τα Collections A , επειδή το μόνο που διαφέρει μεταξύ τους είναι ο αριθμός των παρατηρήσεων , κάθε φορά αυξανόταν για κάθε Collection, τα αντικείμενα που τροποποιηθήκαν κάθε φορά αυξάνονταν και αυτά. Για τα Collections B, στα οποία ο αριθμός των παρατηρήσεων ήταν ίδιος για τα σύνολα δεδομένων B1, B2 και B3 ο αριθμός των αντικειμένων που τροποποιήθηκε μειωνόταν κάθε φορά που αυξανόταν ο αριθμός των ιδιοτήτων που χρησιμοποιήθηκαν. Στα Collections C1, C2, C3 και C4, για τα οποία έγινε συνδυασμός διαφοροποιήσεων μεταξύ των ιδιοτήτων και των αριθμών παρατηρήσεων υπάρχει αύξηση των αντικειμένων που τροποποιήθηκαν όσο αυξανόταν ο αριθμός των παρατηρήσεων.

Όσο αφορά τα Collections D1, D2, D3 και E1, E2, E3 βρίσκονται στο ίδιο σημείο όλα επειδή χρησιμοποιήθηκε ίδιος αριθμός παρατηρήσεων αλλά και ίδιος αριθμός ιδιοτήτων. Διαφέρουν στην παράμετρο όμως που διαφοροποιήθηκε για αυτό υπάρχει η διαφορά στον αριθμό αντικειμένων που τροποποιήθηκε ανάμεσα στα Collections. Για τα Collections D1, D2 και D3 το ποσοστό των τροποποιητών που χρησιμοποιήθηκαν ήταν 2% για όλα και για τους δύο τροποποιητές. Αυτό που διαφοροποιόταν ήταν το ποσοστό αναλογίας των διπλοτύπων έναντι των αριθμών παρατηρήσεων. Ενώ για τα Collections E1, E2 και E3 γινόταν ακριβώς το αντίθετο. Δηλαδή, άλλαζε το ποσοστό των τροποποιητών που χρησιμοποιούταν κάθε φορά από 2%, σε %5 και τέλος σε 10%. Το ποσοστό αναλογίας διπλοτύπων έναντι του αριθμού παρατηρήσεων ήταν το ίδιο και για τα τρία Collections, 2%.

Ακόμη κάτι που παρατηρείται από στο διάγραμμα 5.1, είναι ότι όσο αυξάνετε ο αριθμός των ιδιοτήτων που χρησιμοποιηθήκαν, τόσο αυξάνετε και ο αριθμός των αντικειμένων που τροποποιηθήκαν, ασχέτως αν ο αριθμός των παρατηρήσεων έμενε σταθερός ή αυξανόταν. Αυτό παρατηρείται για τα Collections A, B και C. Φυσικά αυτό μπορεί να οφείλεται και σε αλλαγές σε άλλες παραμέτρους, αύξηση στο ποσοστό των διπλοτύπων ή των τροποποιητών.

5.1 Collection A

Τα Collections A1, A2, A3, A4, A5 και A6 δημιουργήθηκαν έτσι ώστε να μπορούν να χρησιμοποιηθούν για να αξιολογηθούν αλγόριθμοι ταύτισης αντικειμένων ως προς την ιδιότητα της ικανότητας τους να ταυτοποιούν αντικείμενα σε δείγματα διαφορετικών αριθμών παρατηρήσεων αλλά με ίδιες ιδιότητες και παραμέτρους. Το κάθε Collection

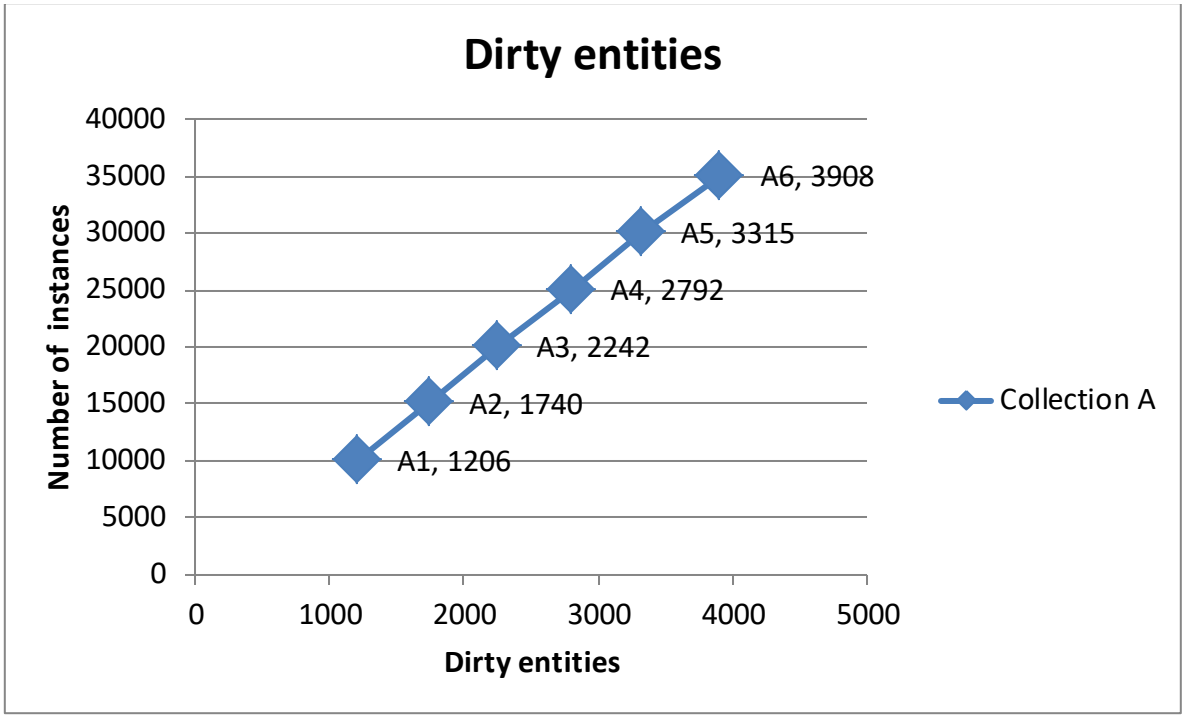
διαφέρει ως προς τον αριθμό των παρατηρήσεων που χρησιμοποιήθηκε. Όσο ο αριθμός των παρατηρήσεων αυξανόταν, τόσο αυξάνονταν και τα αντικείμενα που τροποποιήθηκαν. Στο πίνακα 5.1, παρουσιάζονται αναλυτικά τα αποτελέσματα για την οικογένεια Α. Παρατηρείτε ότι με τις παραμέτρους σταθερές και στα έξι Collections , αλλά κάθε φορά υπήρχε αύξηση του αριθμού των παρατηρήσεων ο αριθμός των τροποποιημένων παρατηρήσεων αυξανόταν. Η αυξητική αυτή τάση φαίνεται στην στήλη «Number of dirty instances» του πίνακα 5.1. Στην στήλη «number of clean instances» του ίδιου πίνακα είναι οι παρατηρήσεις που δεν τροποποιήθηκαν.

Collection A Data sets	Number of dirty entities	Number of clean entities	Total Number of Instances	Ratio Duplicates vs total	Misspelling %	Number of attributes
A1	1206	8794	10000	9%	6%	11
A2	1740	13260	15000	9%	6%	11
A3	2242	17758	20000	9%	6%	11
A4	2792	22208	25000	9%	6%	11
A5	3315	26685	30000	9%	6%	11
A6	3908	31092	35000	9%	6%	11

Πίνακας 5.1: Αποτελέσματα Collection A

Για τα Collections της οικογένειας Α, χρησιμοποιήθηκε ο τροποποιητής για ορθογραφικά λάθη. Οπόταν ο αριθμός των τροποποιημένων παρατηρήσεων είναι το σύνολο των παρατηρήσεων που τροποποιήθηκαν ώστε έτσι να παρουσιάζουν κάποιο ορθογραφικό λάθος.

Στο διάγραμμα 5,2, φαίνεται για κάθε Collection της οικογένειας Α, ο αριθμός των αντικειμένων που τροποποιήθηκε. Παρατηρείτε ότι χωρίς να διαφοροποιείται καμία ιδιότητα μεταξύ των collections, παρά μόνο ο αριθμός των παρατηρήσεων που αυξανόταν κάθε φορά, υπάρχει αύξηση και στα αντικείμενα που τροποποιήθηκαν.



Διάγραμμα 5.2:Collecetion A dirty instances vs clean

Με βάση τα πιο πάνω αποτελέσματα που παρουσιάστηκαν κατά την αξιολόγηση αλγορίθμων ταύτισης αντικειμένων, θα εξεταστεί η ικανότητα του αλγόριθμου να ταυτοποιεί σύνολα δεδομένα που περιέχουν ορθογραφικά λάθη, σε σχετικά μεγάλο αριθμών αντικειμένων. Ως επίσης θα αξιολογείτε και ο χρόνος που χρειάζεται ένας αλγόριθμος να ταυτοποιήσει αντικείμενα που περιέχουν ορθογραφικά λάθη. Όσο αυξάνεται το δείγμα, τόσο ποιο αντιπροσωπευτικά θα είναι και τα αποτέλεσμα της αξιολόγησης ενός αλγορίθμου.

5.2 Collection B

Για κάθε ένα Collection της οικογένειας B, επιλέχθηκε να υπάρχουν κάποιες μετατροπές στις παραμέτρους ώστε να δημιουργηθούν σύνολα από δεδομένα που να σχετίζονται ναι μεν με τις ίδιες ιδιότητες αλλά διαφοροποιημένα και για μεγαλύτερο αριθμό παρατηρήσεων. Για όλα τα Collections που δημιουργηθήκαν η μόνη παράμετρος που παρέμεινε σταθερή ήταν ο αριθμός των παρατηρήσεων που χρησιμοποιήθηκε, 34000. Οι υπόλοιπες παράμετροι διαφοροποιούνταν για κάθε Collection όπως αυτό φαίνεται το Πίνακα 5.2. Επίσης σημαντική παράμετρος και σε αυτή τη συλλογή είναι οι συντακτικές παραλλαγές. Ο Ioannou et al. δημιούργησαν σύνολα δεδομένων που αφορούσαν ορθογραφικά ή συντακτικές παραλλαγές αλλά για μικρό αριθμό ιδιοτήτων

και δείγματος. Ως αυτού δημιουργήσαμε τα σύνολα που αναφαίνονται στον Πίνακα 5.2 και αναφέρονται και λεπτομερώς πιο κάτω.

Στο Collection B1 ως τροποποιητές χρησιμοποιηθήκαν τα ορθογραφικά λάθη και οι μετατροπές λέξεων με ποσοστά 2% και για τις δύο περιπτώσεις. Επίσης η αναλογία διπλότυπων έναντι του αριθμού των αντικειμένων είναι 6%. Για το Collection B1 χρησιμοποιήθηκε αριθμός ιδιοτήτων οχτώ. Στο Collection B2 χρησιμοποιήθηκαν οι τροποποιητές ορθογραφικά λάθη και μετατροπές λέξεων με ποσοστό 5% έκαστος. και αναλογία διπλότυπων έναντι του συνόλου των αντικειμένων 8%. Επίσης αυξήθηκαν οι ιδιότητες σε εννέα. Για την δημιουργία του Collection B3 αυξήθηκε η αναλογία διπλοτύπων έναντι του αριθμού των συνόλων των αντικειμένων σε 10%. Ως ποσοστό των τροποποιητών χρησιμοποιήθηκε 10% και για ορθογραφικά λάθη και μετατροπές λέξεων, και αυξήθηκαν οι ιδιότητες σε 10.

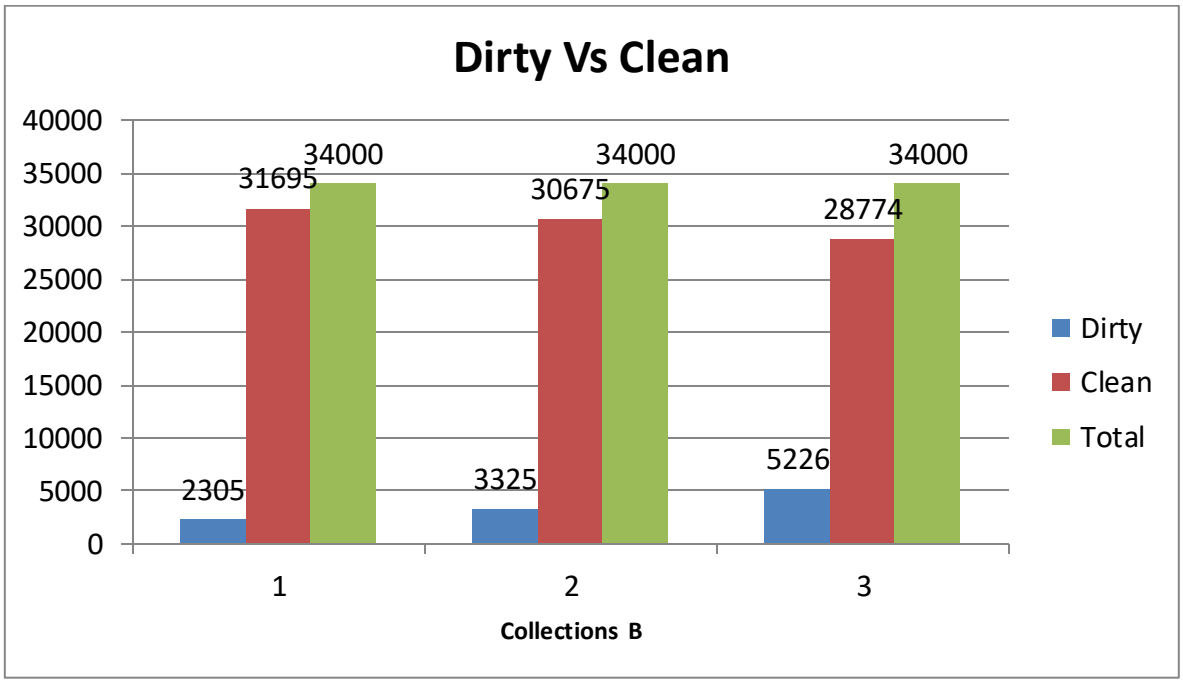
Collection B Data sets	Number of dirty instances	Number of clean instances	Total Number of Instances	Ratio of Duplicates vs total Instances	Misspelling %	Permutation %	Number of attributes
B1	2305	31695	34000	6%	2	2	8
B2	3325	30675	34000	8%	5	5	9
B3	5226	28774	34000	10%	10	10	10

Πίνακας 5.2: Αποτελέσματα Collection B

Στον Πίνακα 5.2 εμφανίζονται τα αποτελέσματα της Collection B. Παρατηρείται πως καθώς αυξάνονταν ιδιότητες που χρησιμοποιηθήκαν, αλλά και τα ποσοστά των τροποποιητών και της αναλογία διπλοτύπων, αυξανόταν παράλληλα και ο αριθμός των αντικειμένων που τροποποιηθήκαν και μειωνόταν παράλληλα ο αριθμός αυτών που δεν τροποποιηθήκαν. Για το Collection B1 το αποτέλεσμα ήταν 2305 τροποποιημένες παρατηρήσεις, για το Collection B2 3325 και για το Collection B3 5226.

Μέσα από τα αποτελέσματα της Collection B (B1, B2, B3) συμπεραίνεται πως (B1, B2, B3) η δημιουργία της μπορεί να αξιολογήσει την ιδιότητα και την συμπεριφορά αλγορίθμων ταύτισης αντικειμένων, να ταυτοποιούν αντικείμενα σε σύνολα δεδομένων που έχουν τον ίδιο αριθμό παρατηρήσεων αλλά διαφέρουν στις υπόλοιπες

παραμέτρους. Επίσης θα μπορεί να αξιολογείται η ποιότητα ενός αλγορίθμου αλλά και ο χρόνος που χρειάζεται ένας αλγόριθμος να ταυτοποιήσει αντικείμενα κάθε φορά σε πιο πολύπλοκο Collection.



Διάγραμμα 5.3: Collection B dirty vs clean

Η αύξηση που αναφέρθηκε πιο πάνω για τα αντικείμενα που τροποποιούνταν κάθε φορά παρουσιάζεται και στο διάγραμμα 5.3. Όπως φαίνεται για σταθερό αριθμό παρατηρήσεων στις 34000, αλλά με διαφορετικές παραμέτρους, κάθε φορά υπήρχε αύξηση για κάθε νέο collection για τις ιδιότητες και τα ποσοστά τρων τροποποιητών. Επιπλέον υπήρχε αύξηση της αναλογίας διπλοτύπων έναντι του συνόλου των παρατηρήσεων, αύξηση των αντικείμενα που τροποποιηθήκαν και μείωση, ως λογικό αποτέλεσμα, αυτών που δεν τροποποιηθήκαν.

5.3 Collection C

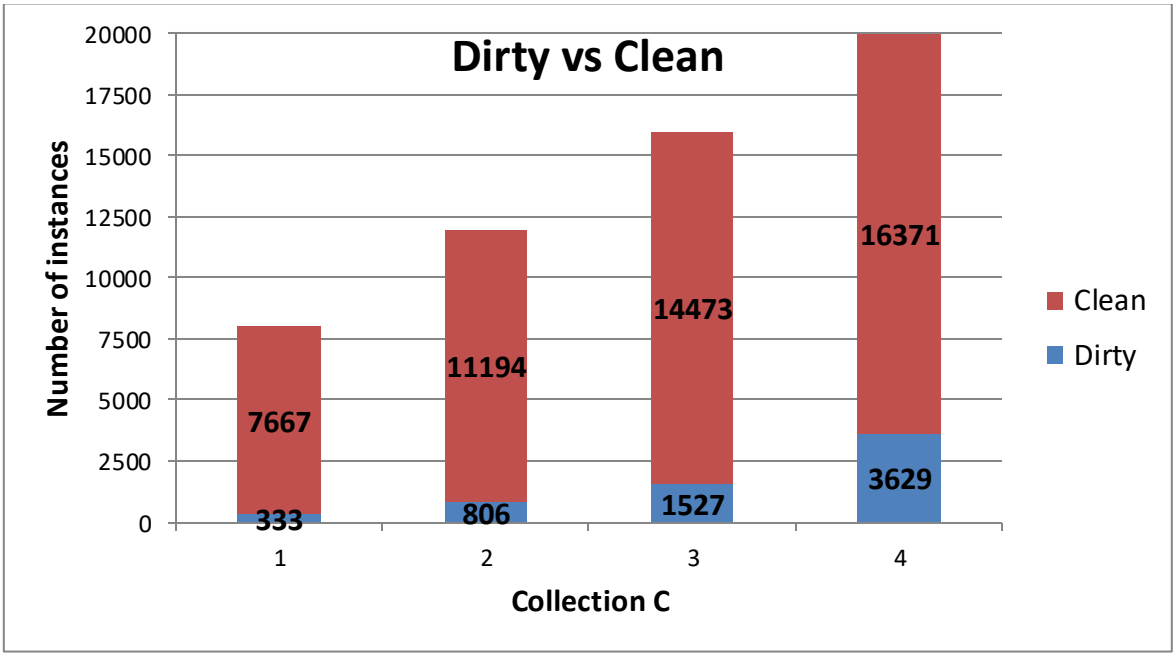
Στο Collection C διαφοροποιούνται όλες οι παράμετροι. Για τα Collections C1, C2, C3 και C4 υπήρξε συνδυασμός διαφοροποιήσεων για όλες τις παραμέτρους για αυτό είναι και η πιο πολύπλοκη οικογένεια. Για κάθε νέο Collection αυξανόταν ο αριθμός των παρατηρήσεων, τα ποσοστά των τροποποιητών που χρησιμοποιηθήκαν και το ποσοστό της αναλογίας διπλότυπων έναντι του συνόλου των παρατηρήσεων. Συγκεκριμένα όπως παρουσιάζεται και στο Πίνακα 5.3, ο αριθμός των παρατηρήσεων ήταν 8000,12000,16000, και 20000 αντίστοιχα για τα Collections C1,C2,C3 και C4. Κάθε

φορά αυξανόταν ο αριθμός των ιδιοτήτων από επτά σε οχτώ για την δημιουργία του C2, σε εννέα για την δημιουργία του C3 και σε δέκα για την δημιουργία του C4. Ενδεικτικά στο Collection C2 προστέθηκε η ιδιότητα publication paper, στο C3 η ιδιότητα publication journal και στο C4 η ιδιότητα publication series.

Collection C Data sets	Number of dirty entities	Number of clean entities	Total Number of instances	Ratio of Duplicates vs total	Acronym %	Permutation %	Number of attributes
C1	333	7667	8000	4%	4	4	7
C2	806	11194	12000	6%	6	6	8
C3	1527	14473	16000	8%	8	8	9
C4	3629	16371	20000	10%	10	10	10

Πίνακα 5.3: Αποτελέσματα Collection C

Ο πίνακας 5.3 παρουσιάζει τον αριθμό των τροποποιημένων παρατηρήσεων για το C1 όπου είναι 333, για το C2 806, για το C3 1527 και για το C4 3629. Για τα Collections αυτά χρησιμοποιήθηκαν οι τροποποιητές μετατροπών λέξεων και ακρωνύμων.



Διάγραμμα 5.4: Collection C dirty vs. Clean

Χαρακτηριστικό της δημιουργίας των collections C1, C2, C3 και C4 είναι καθώς υπάρχει αύξηση στις ιδιότητες και τα ποσοστά των παραμέτρων, τόσο αυξάνονται τα αντικείμενα που τροποποιηθήκαν και μειώνονται αυτά που δεν τροποποιηθήκαν όπως φαίνεται και στο διάγραμμα 5.4. Υπάρχει μια σχέση θετικής συσχέτισης ανάμεσα στην αύξηση των ιδιοτήτων και τα ποσοστά των τροποποιητών και της αναλογίας διπλοτύπων έναντι του συνόλου των παρατηρήσεων με την αύξηση των αντικειμένων που τροποποιήθηκαν.

Τα σύνολα δεδομένων του Collection C που έχουν δημιουργηθεί θα μπορούν να χρησιμοποιηθούν για την αξιολόγηση ενός αλγορίθμου ως προς την ικανότητα του να ταυτοποιεί σύνολα δεδομένων που περιέχουν μεγάλο αριθμό παρατηρήσεων με συντακτικές παραλλαγές. Επίσης θα μπορεί να αξιολογηθεί η συμπεριφορά ενός αλγόριθμου κατά την διαδικασία ταυτοποίησης εξαιτίας της αύξησης και τροποποίησης όλων των παραμέτρων. Ως απόρροια των πιο πάνω ένας ερευνητής θα είναι ικανός να κατανοήσει αν ένας αλγόριθμος ταυτοποιεί σωστά σύνολα δεδομένων που περιέχουν συντακτικά λάθη με μεγάλο αριθμό παρατηρήσεων και αν ένας αλγόριθμος χρειάζεται περισσότερο χρόνο ή όχι να ταυτοποιήσει όταν αυξάνεται ο αριθμός των παρατηρήσεων.

Επίσης, θα μπορεί να αξιολογηθεί και η ποιότητα αλλά και η ικανότητα ενός αλγορίθμου κατά την διαδικασία ταυτοποίησης. Αυτό συμβαίνει λόγω του ότι τα σύνολα περιέχουν συντακτικά λάθη που αυξάνονται ανάμεσα στα σύνολα δεδομένων. Το αποτέλεσμα της αξιολόγησης θα είναι αν ένας αλγόριθμος συμπεριφέρεται διαφορετικά όσο αυξάνεται ο αριθμός των τροποποιημένων παρατηρήσεων.

5.4 Collection D

Για την δημιουργία των collection D1, D2, D3 χρησιμοποιήθηκαν οι ίδιες ιδιότητες του collection B. Ο αριθμός των παρατηρήσεων είναι ίδιος στις 34000 αλλά σε αυτή τη περίπτωση άλλαζει ο αριθμός των ιδιοτήτων σε πέντε για όλα τα collections της οικογένειας D. Το ποσοστό των τροποποιητών παρέμεινε ίδιο και για τα τρία collections, 2%. Το μόνο που διαφοροποιόταν κάθε φορά ήταν το ποσοστό αναλογίας διπλοτύπων έναντι του συνολικού αριθμού παρατηρήσεων. Τα αποτελέσματα καθώς

και οι παράμετροι φαίνονται στο Πίνακα 5.4. Η συλλογή αυτή έχει δημιουργηθεί με σκοπό την εύρεση προβλήματος στα σύνολα οικογενειών αλλά και κατά τη διάρκεια αξιολόγησης του αλγορίθμου.

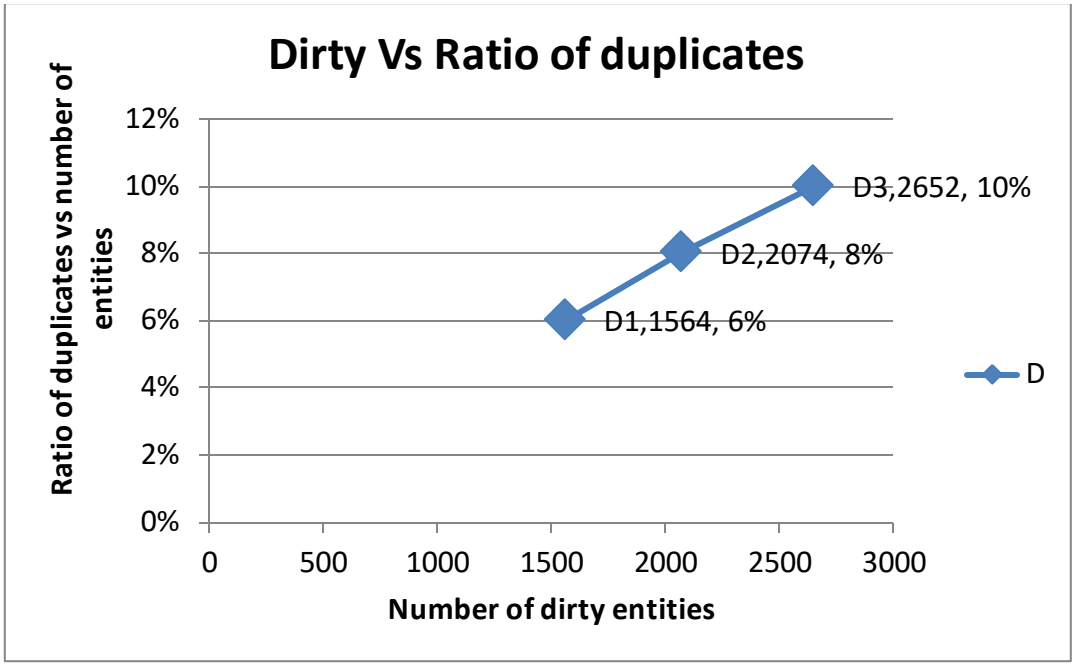
Collection C Data sets	Number of dirty entities	Number of clean entities	Total Number of instances	Ratio of Duplicates vs. total	misspelling %	Permutation %	Number of attributes
D1	1564	32436	34000	6%	2	2	5
D2	2074	31926	34000	8%	2	2	5
D3	2652	31348	34000	10%	2	2	5

Πίνακας 5.4: Αποτελέσματα Collection D

Για ίδιο αριθμό παρατηρήσεων, ίδιο αριθμών ιδιοτήτων, ίδιο ποσοστό τροποποιητών αλλά διαφορετικό ποσοστό της αναλογίας των διπλοτύπων έναντι του συνολικού αριθμού παρατηρήσεων παρατηρήθηκε αύξηση στον αριθμών αντικειμένων που τροποποιήθηκαν σε κάθε collection.

Με βάση τα αποτελέσματα της οικογένειας D, παρατηρείται πως τα σύνολα δεδομένων που δημιουργούνται μπορούν να χρησιμοποιηθούν για τον εντοπισμό προβλήματος κατά την αξιολόγηση ενός αλγορίθμου ταύτισης αντικειμένων. Δηλαδή, εάν ένας αλγόριθμος χρησιμοποιηθεί σε πιο πολύπλοκα σύνολα δεδομένων όπως της οικογένειας B και C, και παρουσιάσει κάποιο πρόβλημα κατά την ταυτοποίηση, τότε μπορούν να χρησιμοποιηθούν τα σύνολα δεδομένα της οικογένειας D, ώστε εάν παρουσιάσει ξανά πρόβλημα, τότε θα οφείλεται στο ποσοστό της αναλογίας των διπλοτύπων έναντι του συνολικού αριθμού παρατηρήσεων, γιατί είναι η μόνη παράμετρος που διαφοροποιείται ανάμεσα στα σύνολα δεδομένων της οικογένειας D.

Στο διάγραμμα 5.5 φαίνεται και παραστατικά η θετική σχέση συσχέτισης μεταξύ της αναλογίας διπλοτύπων έναντι του συνολικού αριθμού παρατηρήσεων με τα αντικείμενα που έχουν τροποποιηθεί. Καθώς αυξάνεται το ποσοστό αυξάνεται και ο αριθμός των αντικειμένων που τροποποιηθήκαν με τις υπόλοιπες παραμέτρους και των αριθμών των παρατηρήσεων να μην διαφοροποιούνται μεταξύ των συνόλων δεδομένων D1,D2 και D3.



Διάγραμμα 5.5: Collection D dirty vs. Ration of duplicates

5.5 Collection E

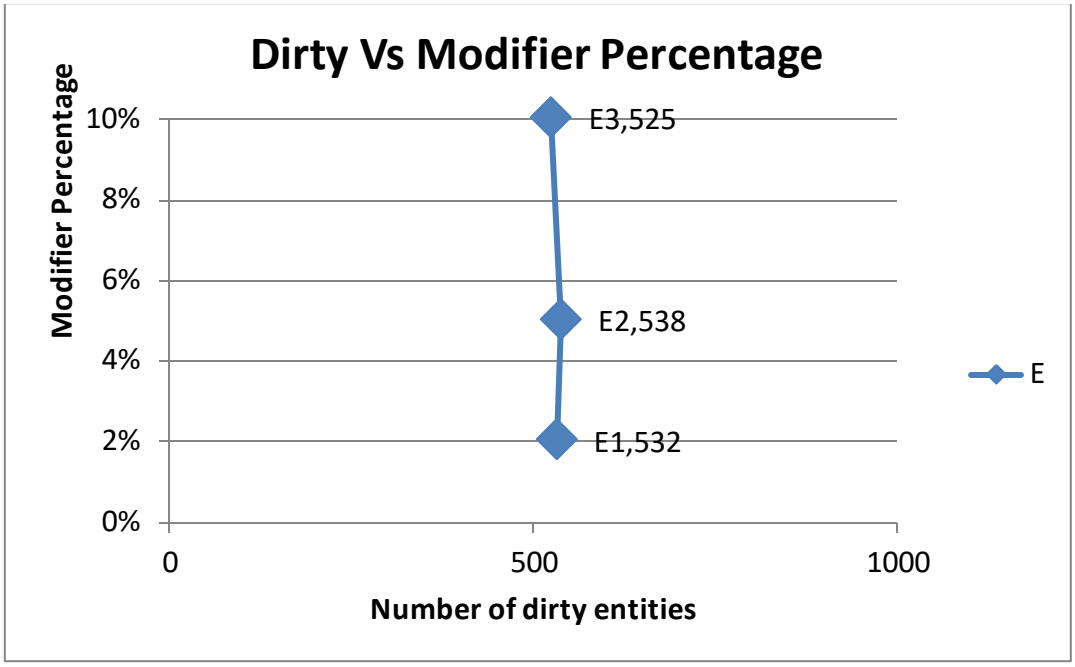
Για την δημιουργία των collection E1, E2 και E3 χρησιμοποιήθηκαν όπως και στα Collections D1, D2 και D3 οι ίδιες ιδιότητες του collection B. Ο αριθμός των παρατηρήσεων είναι ίδιος στις 34000 αλλά άλλαξε ο αριθμός των ιδιοτήτων σε πέντε για όλα τα collections της οικογένειας D. Σε αντίθεση όμως με τα σύνολα δεδομένων της οικογένειας D, αυτή την φορά το μόνο που διαφοροποιόταν κάθε φορά ήταν το ποσοστό των τροποποιητών που χρησιμοποιηθήκαν. Το ποσοστό αναλογίας διπλοτύπων έναντι του αριθμού των συνολικών παρατηρήσεων παρέμεινε ίδιο και για τα τρία collections, στο 2% Τα αποτελέσματα καθώς και οι παράμετροι φαίνονται στο Πίνακα 5.5. Το Collection E, όπως και στη περίπτωση του D, έχει δημιουργηθεί για την εύρεση προβλήματος όταν μεταβάλλεται μόνο το ποσοστό του τροποποιητή.

Collection C Data sets	Number of dirty entities	Number of clean entities	Total Number of instances	Ratio of Duplicates vs. total	misspelling %	Permutation %	Number of attributes
E1	532	33468	34000	2%	2	2	5
E2	538	33462	34000	2%	5	5	5
E3	525	33475	34000	2%	10	10	5

Πίνακας 5.5: Αποτελέσματα Collection E

Παρατηρείται ότι καθώς αυξάνεται το ποσοστό των τροποποιητών που χρησιμοποιηθήκαν , ενώ ο αριθμός των παρατηρήσεων παραμένει ο ίδιος καθώς και ούτε οι υπόλοιπες παράμετροι διαφοροποιούνται, δεν υπάρχει συσχέτιση μεταξύ των αντικειμένων που τροποποιήθηκαν κάθε φορά. Ο χαμηλός αριθμός των αντικειμένων που τροποποιήθηκε σε σχέση με τον αριθμό των παρατηρήσεων που είναι στις 34000, οφείλεται στο χαμηλό ποσοστό της αναλογίας των διπλοτύπων έναντι του συνολικού αριθμού παρατηρήσεων, που ήταν 2% και για τα τρία collections της οικογένειας Ε. Οι παραπάνω παρατηρήσεις παρουσιάζονται στο διάγραμμα 5.6.

Σκοπός της δημιουργία των Collection E1, E2 και E3, είναι να υπάρχουν σύνολα δεδομένων που διαφοροποιείται μόνο συγκεκριμένη παράμετρος, στην περίπτωση αυτή το ποσοστό των τροποποιητών, έτσι ώστε να μπορούν να χρησιμοποιηθούν για την εύρεση προβλήματος ενός αλγορίθμου. Δηλαδή, εάν κατά την διάρκεια αξιολόγησης ενός αλγορίθμου σε πιο πολύπλοκα δεδομένα που έχουν περισσότερες διαφοροποιήσεις παρουσιαστεί κάποιο πρόβλημα, μπορούν να χρησιμοποιηθούν τα collections E1, E2 και E2 που διαφέρουν μόνο στο ποσοστό του τροποποιητή, έτσι αν παρουσιαστεί ξανά το συγκεκριμένο πρόβλημα τότε αυτό θα οφείλεται σε πρόβλημα που προκύπτει στα ποσοστά των τροποποιητών η ακόμη και στους ίδιους τους τροποποιητές.



Διάγραμμα 5.6: Collection E dirty vs. Modifier percentage.

5.6 Συζήτηση

Συνοψίζοντας την παραπάνω ανάλυση των αποτελεσμάτων παρατηρούμε πως οι πέντε οικογένειες που δημιουργήθηκαν διαφέρουν ώστε να υπάρχουν σύνολα για χρήση αξιολόγησης αλγορίθμων, που θα μπορούν να χρησιμοποιηθούν αναλόγως της ανάγκης ενός ερευνητή κατά την αξιολόγηση ενός αλγόριθμου ταύτισης αντικειμένων.

Στα σύνολα δεδομένων του Collection A, το μόνο που διαφέρει μεταξύ τους είναι ο αριθμός των παρατηρήσεων ο οποίος σε κάθε νέο σύνολο δεδομένων αυξάνεται. Οπόταν τα σύνολα μπορούν να χρησιμοποιηθούν σε περίπτωση που κάποιος ερευνητής θέλει να αξιολογήσει την συμπεριφορά ενός αλγόριθμου καθώς αυξάνεται ο αριθμός των παρατηρήσεων καθώς και το χρόνο που θα χρειαστεί για την ταυτοποίηση. Στο Collection B, αντίθετα, ο αριθμός των παρατηρήσεων παραμένει σταθερός, αλλά διαφοροποιούνται και αυξάνονται τα ποσοστά και ο αριθμός των υπόλοιπων παραμέτρων. Τα σύνολα δεδομένων του Collection B, μπορούν να χρησιμοποιηθούν για την αξιολόγηση αλγορίθμων ταύτισης αντικειμένων ως προς την ικανότητα τους να ταυτοποιούν σε σύνολα δεδομένων που διαφοροποιούνται όλοι οι παράμετροι εκτός από τον αριθμό αντικειμένων. Τέλος, τα σύνολα δεδομένων του Collection C, θα μπορούν να χρησιμοποιηθούν για την αξιολόγηση ενός αλγορίθμου ως προς την

συμπεριφορά αυτού όσο αυξάνεται και ο αριθμός των παρατηρήσεων αλλά και διαφοροποιούνται όλοι οι παράμετροι κάθε φορά. Τα σύνολα δεδομένων του Collection C μπορούν να προσφέρουν την μέγιστη αξιολόγηση για ένα αλγόριθμο καθώς είναι πιο πολύπλοκα από τα υπόλοιπα.

Στα σύνολα δεδομένων των Collections B και C παρατηρείτε η μη αναγνώριση προβλήματος κατά την διαδικασία αξιολόγησης, εξαιτίας της ταυτόχρονης διαφοροποίησης των παραμέτρων κάθε φορά. Σε αντίθεση στο Collection A δεν παρατηρείται αυτό το φαινόμενο λόγω της διαφοροποίησης μόνο μιας παραμέτρου. Ωστόσο, δημιουργηθήκαν τα σύνολα δεδομένων των Collections D και E όπου έχουν ως πλεονέκτημα την εύρεση προβλήματος που εμφανίζεται κατά την αξιολόγηση ενός αλγορίθμου. Αυτό συμβαίνει καθότι στα Collections D και E διαφοροποιείτε μόνο μια παράμετρος, με αποτέλεσμα σε περίπτωση που παρουσιαστεί ξανά το πρόβλημα κατά την ταυτοποίηση χρησιμοποιώντας τα Collections D ή E θα οφείλεται λόγω της συγκεκριμένης παραμέτρου. Επίσης μπορούν να χρησιμοποιηθούν και για την αξιολόγηση ενός αλγορίθμου ταύτισης αντικειμένων.

Κεφάλαιο 6

Επίλογος

Ο στόχος της παρούσας μεταπτυχιακής διατριβής ήταν ο καθορισμός οικογενειών συνόλων δεδομένων βάση υφιστάμενων βιβλιογραφικών ερευνών ως και επίσης η δημιουργία αυτών για χρήση αξιολόγησης αλγορίθμων ταύτισης αντικειμένων. Στο κεφάλαιο 5 αναλύθηκαν και συζητήθηκαν διαφορετικές καινοτόμες οικογένειες συνόλων που θα μπορούσαν να χρησιμοποιηθούν στους διάφορους αλγόριθμους για επίλυση του υφιστάμενου προβλήματος.

Στην εν λόγω μεταπτυχιακή διατριβή δημιουργήθηκαν καινοτόμα σύνολα δεδομένων που αφορούν δημοσιεύσεις, ταινίες και ερευνητές. Τα σύνολα δεδομένων που δημιουργήθηκαν με σκοπό την αξιολόγηση αλγορίθμων ταύτισης αντικειμένων διαφέρουν αρκετά σε παραμέτρους και ιδιότητες ώστε να υπάρχουν διάφορων μορφής σύνολα δεδομένων. Υπάρχουν πολλοί συνδυασμοί διαφοροποιήσεων των παραμέτρων που μπορούν να γίνουν για την δημιουργία συνόλων δεδομένων. Αρχικά δημιουργήθηκαν πολύπλοκες οικογένειες συνόλων έτσι ώστε να γίνει καλύτερη αξιολόγηση των αλγορίθμων ως προς την ποιότητα, το χρόνο που χρειάζεται για την ταυτοποίηση αντικειμένων και την συμπεριφορά τους κατά την ταυτοποίηση(scalability). Με βάση τα Collections που αναλύθηκαν στο προηγούμενο κεφάλαιο η συλλογή στην οποία τροποποιούνται όλες οι παράμετροι δημιουργεί τις καταλληλότερες προϋποθέσεις για τη μέγιστη αξιολόγηση.

Τα σύνολα αυτά δίνουν τη δυνατότητα εισαγωγής μεγαλύτερου αριθμού δείγματος και παράλληλα πρόσθετων ιδιοτήτων με αποτέλεσμα μεγαλύτερης πιθανότητας ταύτισης των αντικειμένων. Επιπρόσθετα τα σύνολα δεδομένων που δημιουργήθηκαν επιτρέπουν την εύρεση λάθους κάτι που στα προηγούμενα σύνολα δεδομένων από προηγούμενες μελέτες υπήρχε ως περιορισμός. Περεταίρω τα σύνολα αυτά θα μπορούν να αξιολογηθούν από οποιοδήποτε ερευνητή οποίος θα τα χρησιμοποιήσει για αξιολόγηση των αλγορίθμων.

Η δημιουργία των οικογενειών συνόλων δεδομένων έγινε χρησιμοποιώντας το σύστημα EMBench. Το σύστημα EMBench θεωρήθηκε ως το καταλληλότερο για να

δημιουργηθούν οι οικογένειες γιατί έχει ως πλεονέκτημα σε σχέση με άλλα συστήματα που αναφέρθηκαν στην βιβλιογραφική επισκόπηση τη δυνατότητα πλήρης τροποποίησης των παραμέτρων από τον χρήστη.

Για μελλοντική έρευνα, ερευνητές και συγγραφείς από την ερευνητική κοινότητα, μπορούν με βάση τις ιδιότητες των οικογενειών των συνόλων δεδομένων που δημιουργηθήκαν και χρησιμοποιώντας το σύστημα EMBench ή κάποιο άλλο σύστημα που να προσφέρει την δυνατότητα διαφοροποιήσεων των παραμέτρων, να δημιουργήσουν νέα σύνολα δεδομένα διαφοροποιημένα ως προς τις παρατηρήσεις, τις ιδιότητες και τις παραμέτρους. Επίσης θα μπορούσαν, να δημιουργήσουν σύνολα δεδομένα που να περιέχουν παρατηρήσεις από αυτές που δημιουργηθήκαν στο προηγούμενο στάδιο (σύνολο δεδομένων) και όχι από τα αρχικές πηγές. Τέλος θα μπορούσαν να χρησιμοποιούσουν τα σύνολα δεδομένων της παρούσας μεταπτυχιακής διατριβής για αξιολόγηση αλγορίθμων ταύτισης αντικειμένων.

Βιβλιογραφία

[BG07] I. Bhattacharya; L. Getoor, "Collective Entity Resolution in Relational Data," in ACM Transactions on Knowledge Discovery From Data, 2007, vol. 1, article 5.

[BGM+09] O. Benjelloun; H. Garcia-Molina; D. Menestrina; Q. Su; S. Whang; J. Widom , "Swoosh:A Generic Approach to Entity Resolution," in International Journal on Very Large Data Bases, 2009, vol. 18, No.1, pp.255-276.

[CJZ+12] J. Chen; C. Jin; R. Zhang; R. A. Zhou, "A Learning Method for Entity Matching," in 10th International Workshop on Quality in Databases, 2012.

[DHM05] X. Dong; A. Halevy; J. Madhavan, "Reference Reconciliation in Complex Information Spaces," in ACM SIGMOD international conference on Management of data, 2005, pp. 85-96.

[FMN+11] A. Ferrara; S. Montanelli; J. Noessner; H. Stuckenschmidt, "Benchmarking Matching Applications on the Semantic Web," in 8th extended semantic web conference on The semantic web: research and applications ,2011, vol. 6644, pp. 108-122.

[GB06] L. Gu; R. Baxter, "Decision Models for Record Linkage," Data Mining, LNAI 3755, pp. 146-160, 2006.

[HKB14] S. Homoceanu; J-C. Kalo; W-T. Balke, "Putting Instance Matching to the Test: Is Instance Matching Ready for Reliable Data Linking?," in 21st International Symposium on Methodologies for Intelligent Systems (ISMIS) , 2014, pp. 274-284.

[HT12] D. M. Herzig; T. Tran, "Heterogeneous Web Data Search Using Relevance-based On The Fly Data Integration," in 21st international conference on World Wide Web, 2012, pp. 141-150.

[IV14] E. Ioannou; Y. Velegrakis, "EMBench: Generating Entity-Related Benchmark Data," in International Conference on Posters & Demonstrations Track, 2014, vol. 1272, pp. 113-116.

[IRV13] E. Ioannou; N. Rassadko; Y. Velegrakis, “On Generating Benchmark Data for Entity Matching,” Journal of Data Semantics, vol. 2, pp. 37-56, 2013.

[KR10] H. Kopcke; E. Rahm, “Frameworks for Entity Matching: A comparison,” Data & Knowledge Engineering, vol. 69, pp. 197-210, 2010.

[KTR09] H. Kopcke; A. Thor; E. Rahm, “Comparative Evaluation of Entity Resolution Approaches with Fever,” in Comparative Evaluation of Entity Resolution Approaches with Fever, 2009, vol. 2, pp. 1574-1577.

[KTR10] H. Kopcke; A. Thor; E. Rahm, “Evaluation of Learning-Based approaches for Matching Web Data Entities,” IEEE Internet Computing, vol. 14, no. 4, pp. 23-31, 2010.

[LLH14] S. Lee; J. Lee ; S-W. Hwang, “Efficient Entity Matching Using Materializes Lists,” Information Sciences: an International Journal, vol. 261, pp. 170-184, 2014.

[MBB+10] Z. Miklos; N. Bonvin; P. Bouquet; M. Catasta; D. Cordioli; P. Fankhauser; J. Gaugaz; E. Ioannou; H. Koshutanski; A. Mana; C. Niederee; T. Palpanas; H. Stoermer,” From Web Data to Entities and Back,” in 22nd international conference on Advanced information systems engineering, 2010, vol. 6051, pp. 302-316.

[MDR+12] R. Mirrizi; T. Di Noia; A. Ragone; V. C. Ostuni; E. Di Sciascio, “Movie recommendation with Dbpedia,” in 3rd Italian Information Retrieval Workshop, 2012.

[MSM16] S. Mishra; S. Saha; S. Mondal, “A Multiobjective Optimization Based Entity Matching Technique for Bibliographic Databases,” Expert Systems With Applications, vol. 65, pp. 100-115, 2016.

[PIN+12] G. Papadakis; E. Ioannou; C. Niederee; T. Palpanas; W. Nejdl, “Beyond 100 Million Entities: Large-Scale Blocking-Based Resolution for Heterogeneous Data,” in fifth ACM international conference on Web search and data mining, 2012, pp. 53-62.

[SB02] S. Sarawagi; A. Bhamidipaty, “Interactive Deduplication using Active Learning,” in eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002, pp. 269-278.

[SBK+02] S. Sarawagi; A. Bhamidipaty; A. Kirpal; C. Mouli, “ALIAS: An Active Learning led Interactive Deduplication System,” in 28th international conference on Very Large Data Bases, 2002, pp. 1103-1106.

[SK03] S. Sarawagi; A. Kirpal, “Scaling up the ALIAS Duplicate Elimination System: A Demonstration,” in 19th International Conference on Data Engineering, 2003, pp. 783-785.

[SDV+07] W. Shen; P. DeRose; L. Vu; A. Doan; R. Ramakrishnan, “Source-aware Entity Matching: A Compositional Approach,” in IEEE 23rd International Conference on Data Engineering, 2007, pp. 196-205.

[TZY+08] J. Tang; J. Zhang; L. Yao; J. Li; L. Zhang; Z. Su, “ArnetMiner: extraction and mining of academic social networks,” in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008, pp. 990-998.s

[YHY07] X. Yin; J. Han; P.S. Yu, “Object Distinction: Distinquishing Objects with Identical Names,” in Distinquishing Objects with Identical Names”, Data Engineering, 2007, pp. 1242-1246.