Ατομική Διπλωματική Εργασία

A heuristic algorithm for matching user Online Social Network profiles

Κυριάκος Φραγκέσκος

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάιος 2015

Πανεπιστήμιο Κύπρου
Τμήμα Πληροφορικής

Develop a heuristic algorithm to perform matching between the users in a variety of social networks

Κυριάκος Φραγκέσκος

Επιβλέπων Καθηγητής
Δρ. Γιώργος Πάλλης

Η Ατομική Διπλωματική Εργασία υποβλήθηκε προς μερική εκπλήρωση των απαιτήσεων απόκτησης του πτυχίου Πληροφορικής του Τμήματος Πληροφορικής του Πανεπιστημίου Κύπρου

Μάιος 2015

Στους γονείς μου, Ανδρέα και Σταύρη Φραγκέσκου.
Στην Κατερίνα Παντελή.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου Δρ. Γιώργο Πάλλη, για την βοήθεια του και για την πολύ καλή συνεργασίας που είχαμε στα πλαίσια της διπλωματικής μου εργασίας. Οι συμβουλές και η καθοδήγηση ήταν αρωγοί στο να ολοκληρώσω, με επιτυχία κατά την άποψη μου, την διπλωματική μου εργασία.

Στη συνέχεια θα ήθελα να εκφράσω την ευγνωμοσύνη μου και τις ευχαριστίες μου στους Δρ. Δημήτρη Αντωνιάδη και υποψήφιο διδάκτορα Χαρίτων Ευσταθιάδη, για την πολύτιμη υποστήριξη που μου πρόσφεραν, καθώς επίσης και για τις πολύτιμες συμβουλές και το freedback που μου έδιναν όλο αυτό το διάστημα με σκοπό να βελτιώσω, στο χρόνο που είχα, στο μέγιστο την διπλωματική εργασία μου.

Τέλος, θα ήθελα να εκφράσω ιδιαίτερες ευχαριστίες στην οικογένεια μου που με στήριξαν ηθικά, ψυχολογικά και οικονομικά σε ολόκληρη την πορεία μου κατά τη διάρκεια φοίτησης μου στο Πανεπιστήμιο Κύπρου, αλλά και κατά τη δύσκολη αυτή περίοδο ολοκλήρωσης της διπλωματικής μου εργασίας.

Περίληψη

Ζούμε σε ένα κόσμο όπου η συνεχής ανάπτυξη της τεχνολογίας είναι πλέον δεδομένη. Θα πρέπει να ακολουθούμε αυτούς τούς ρυθμούς για να μπορέσουμε να συλλάβουμε και να κατανοήσουμε τα τεχνολογικά επιτεύγματα τα οποία συντελούνται καθημερινά. Σε αυτό τον τεχνολογικά προηγμένο κόσμο υπάρχουν και τα κοινωνικά δίκτυα όπου μας ανοίγουν ένα νέο παράθυρο ενημέρωσης και επικοινωνίας.

Σε αυτή την διπλωματική εργασία, προσπαθήσαμε να εισηγηθούμε έναν ευρετικό τρόπο, που να είναι σε θέση να βρίσκει όλα τα πιθανά προφίλ που έχει ένας χρήστης στα κοινωνικά δίκτυα. Πιο αναλυτικά, δοθέντος ενός profile url για οποιοδήποτε προφίλ κάποιου χρήστη, συλλέγουμε όλες τις διαθέσιμες πληροφορίες για τον συγκεκριμένο χρήστη και εφαρμόζουμε τον αλγόριθμο οποίος πηγαίνει και βρίσκει τα προφίλ στα υπόλοιπα κοινωνικά δίκτυα τα οποία πιθανόν να ανήκουν στον ίδιο χρήστη. Η προσέγγισή μας βγάζει μία πιθανότητα που μεταφράζεται σε ποσοστό ακρίβειας όσο αφορά την ταύτιση των προφίλ.

Στα πλαίσια της διπλωματικής εργασίας ασχοληθήκαμε με τους συνδυασμούς Twitter - LinkedIn και Facebook - Flickr. Δηλαδή, υλοποιήσαμε και εξετάσαμε κατά πόσο είναι εφικτό, κάνοντας εφαρμογή της μεθόδου μας, να εξάγουμε το Twitter account κάποιου χρήστη χρησιμοποιώντας το LinkedIn account του, και αντίστροφα. Παράλληλα εξετάσαμε και την περίπτωση να εξάγουμε το Facebook account κάποιου χρήστη κάνοντας χρήση του Flickr account του.

Μετέπειτα, αφού αξιολογήσαμε την προσέγγιση μας με τετριμμένες προσεγγίσεις, προχωρήσαμε στην υλοποίηση ενός εργαλείο που να επιτρέπει σε κάποιον χρήστη να χρησιμοποιεί την λειτουργία της ταύτισης προφίλ. Το εργαλείο αυτό είναι μία διαδικτυακή εφαρμογή με το όνομα Wally υλοποιημένο με τεχνολογία Mean Stack. Η διαδικτυακή εφαρμογή εφαρμόζει τον αλγόριθμο που εισηγούμαστε.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Σε αυτό το κεφάλαιο θα παρουσιαστούν κάποια προβλήματα τα οποία ταλανίζουν το ευρύ κοινό και τα οποία έγιναν αφορμή να ασχοληθούμε με τα κοινωνικά δίκτυα και τα big data.

1.1 Κοινωνικά Δίκτυα

Όπως αναφέρεται στο Wikipedia [16], ένα κοινωνικό δίκτυο είναι μια κοινωνική δομή που αποτελείται από ένα σύνολο κοινωνικών φορέων (όπως τα άτομα ή οργανώσεις) και ένα σύνολο από τους δυαδικούς δεσμούς μεταξύ αυτών των φορέων.

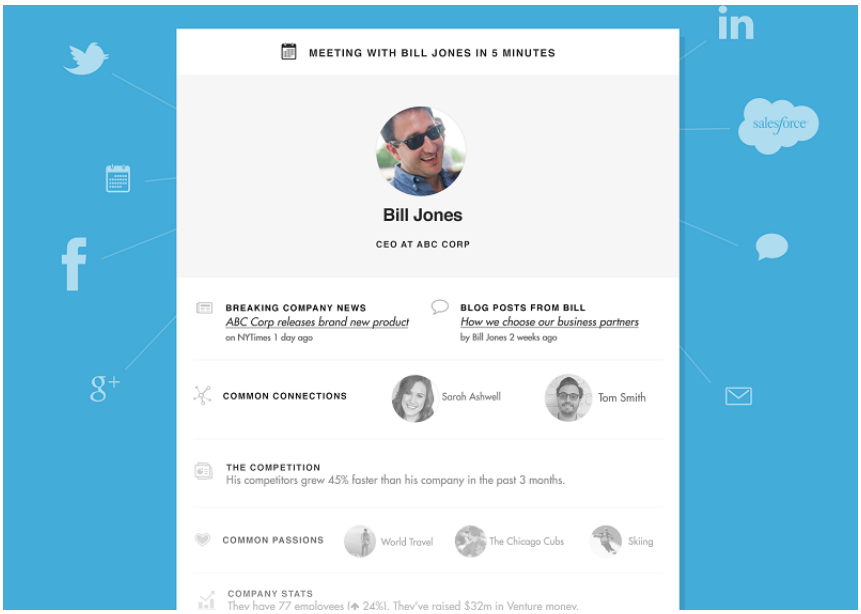
Μερικά παραδείγματα κοινωνικών δικτύων είναι, το Facebook όπου είναι ιστοχώρος κοινωνικής δικτύωσης που ξεκίνησε στις 4 Φεβρουαρίου του 2004. Οι χρήστες μπορούν να επικοινωνούν μέσω μηνυμάτων με τις επαφές τους και να τους ειδοποιούν όταν ανανεώνουν τις προσωπικές πληροφορίες τους.

Το Flickr είναι μια πλατφόρμα όπου οι χρήστες έχουν την δυνατότητα να ανεβάζουν φωτογραφίες τους. Το Flickr παρέχει στους χρήστες του 1 Terabyte, όπου είναι διαθέσιμο για αποθήκευση φωτογραφιών.

Ένα ανερχόμενο κοινωνικό δίκτυο είναι το Twitter, όπου είναι ένας ιστοχώρος κοινωνικής δικτύωσης που επιτρέπει στους χρήστες του να στέλνουν και να διαβάζουν σύντομα μηνύματα. Τα μηνύματα μπορούν να αναγνωστούν και από μη συνδεδεμένους χρήστες, αλλά μόνο οι συνδεδεμένοι μπορούν να δημοσιεύσουν κείμενα.

Το LinkedIn είναι ένας ιστοχώρος επαγγελματικής κοινωνικής δικτύωσης. Ιδρύθηκε τον Δεκέμβριο του 2002. Τα εγγεγραμμένα μέλη του έχουν τη δυνατότητα να δημιουργήσουν το προσωπικό επαγγελματικό τους προφίλ, να συνδεθούν με άλλους χρήστες, να αναζητήσουν εργασία, αλλά και να δημιουργήσουν πελατολόγιο.

Τα κοινωνικά δίκτυα είναι ένας ενδιαφέρων κόσμος, όπου προσελκύει πάρα πολλούς



Σχήμα 1.1: Charlie's Results

ερευνητές. Για παράδειγμα ο R.Jain and D.Sonnen [2] εισηγούνται ένα νέο είδος κοινωνικού δικτύου. Το οποίο θα επιτρέπει στους χρήστες να δημοσιεύουν σε μορφή micro blogs είτε κείμενο είτε ήχο ή/και εικόνα.

Επίσης, προτείνουν ένα καινοτόμο τρόπο, ώστε η πλειοψηφία των ανθρώπων να μπορούν να χρησιμοποιούν εύκολα αυτά τα δίκτυα. Και με τη χρήση αυτή, να εκμεταλλευτούν και να αποκτήσουν χρήσιμες πληροφορίες από αυτό το κοινωνικό δίκτυο.
Το αφαιρετικό μοντέλο που εισηγούνται ακολουθεί την πιο κάτω δομή: Event >> Sensors >> Aggregation >> Detection >> Info/Alert
Οι κύριες πηγές πληροφόρησης για συλλογή πληροφοριών είναι:

1. Data on the Web
2. Real- time micro-blogs on the web
3. from other sensors (e.g. camera etc)

Και σε συνδυασμό με push/pull base and filtering techniques μπορούν εύκολα να εξάγουν χρήσιμες πληροφορίες. Δημιουργούν ένα σύστημα ενημέρωσης μέσω των κοινωνικών δικτύων όπου θα ενημερώνουν σε πραγματικό χρόνο για events που γίνονται.

Δεν υπάρχουν μόνο μελέτες που ασχολούνται με τα κοινωνικά δίκτυα, υπάρχουν και εφαρμογές που κάνουν χρήση των κοινωνικών δικτύων. Όπως για παράδειγμα ο Charlie [1], που είναι ένα εργαλείο που συγκεντρώνει πληροφορίες για τους ανθρώπους

[1]https://charlieapp.com/

Σχήμα 1.2: Social Media Users Jan 2015

που πρόκειται να συναντήσεις, πριν τους δεις. Μία ώρα πριν τη συνάντηση στο ημερολόγιο σου ο Charlie εξασφαλίζει ότι μπαίνοντας στη συνάντηση θα κάνεις θεαματική εντύπωση διότι θα γνωρίζεις τα πάντα για τα άτομα που θα βρίσκονται στην συνάντηση, όπως για τις ασχολίες τους, το επαγγελματικό τους υπόβαθρο κ.α.(βλέπε Σχ. 1.1)

Υπάρχουν δεκάδες κοινωνικά δίκτυα όπου χωρίζονται σε κατηγορίες ανάλογα με τις λειτουργίες που παρέχουν στους χρήστες. Ως ερευνητές χρησιμοποιούμε τα κοινωνικά δίκτυα επειδή μπορούν να παρέχουν χρήσιμες πληροφορίες ανά πάσα στιγμή. Επίσης ο κόσμος των Social Media συνεχώς αναπτύσσεται, παρέχοντας μεγάλο όγκο δεδομένων. Αξίζει να σημειωθεί ότι μέσα από τα κοινωνικά δίκτυα μπορούμε να εκμεταλλευτούμε τα δεδομένα που παρέχονται για να κάνουμε τη ζωή των ανθρώπων καλύτερη.

1.1.1 Ενδιαφέροντα Στοιχεία

Κάποιες πληροφορίες όσο αφορά κάποια από τα κοινωνικά δίκτυα που αντικατοπτρίζουν το πόσο μεγάλο αντίκτυπο και πόσο πολύ επηρεάζουν την καθημερινή μας ζωή.

1. Υπάρχουν περίπου 1,4 δισεκατομμύρια χρήστες στο Facebook.
2. 47% από όλους του χρήστες στο διαδίκτυο είναι στο Facebook.
3. 4,5 δισεκατομμύρια likes παράγονται καθημερινά.
4. Το Twitter έχει 284 εκατομμύρια χρήστες μέχρις στιγμής.

5. Παράγονται 500 εκατομμύρια tweets καθημερινά.

6. Το LinkedIn έχει 347 εκατομμύρια χρήστες.

7. Το LinkedIn είχε κέρδος $643 εκατομμύρια

Βλέποντας τα πιο πάνω καταλαβαίνει κανείς πόσο συνυφασμένα είναι με την καθημερινότητα μας τα κοινωνικά δίκτυα.[2]

1.2 Πρόβλημα/Motivation

Το κύριο πρόβλημα που θέλουμε να εξαλείψουμε ή να μειώσουμε στο μέγιστο είναι αυτό του χρόνου αναζήτησης καθώς και της ακρίβειας των αποτελεσμάτων όσο αφορά την ταύτιση των διάφορων προφίλ που έχει ένας χρήστης στα κοινωνικά δίκτυα. Συνεπώς, το πρόβλημα είναι ο χρόνος που σπαταλιέται άσκοπα πάνω σε διάφορες αναζητήσεις ατόμων στα κοινωνικά δίκτυα και η ανάγκη και αποδοτικής εύρεσης και εξαγωγής σωστών αποτελεσμάτων.

Εκτός από το κύριο πρόβλημα μας, σκεφτήκαμε και μελλοντικές εργασίες που θα μπορούσαν να εκμεταλλευτούν τυχόν λύση του προβλήματος της ταύτισης των προφίλ. Όπως για παράδειγμα η ανάπτυξη ενός έξυπνου recommendation system, που να προτείνει σε έναν χρήστη ένα προϊόν που να συμβαδίζει με τα χαρακτηριστικά του. Για παράδειγμα το Netflix ή Amazon όπου προτείνουν ταινίες και προϊόντα αντίστοιχα, στους χρήστες τους με βάση το ιστορικό τους.

Αυτό το έξυπνο recommendation system θα μπορούσε να αποφύγει κάποια προβλήματα που αντιμετωπίζουν τα περισσότερα recommendation systems όπως το πρόβλημα του cold star. Το cold star είναι το φαινόμενο που παρατηρείται όπου ένα recommendation system στην αρχή δεν έχει τα απαραίτητα στοιχεία για να προτείνει κάποιο προϊών στον χρήστη.

Επίσης ένα άλλο πρόβλημα είναι ότι υπάρχει μεγάλος αριθμός από κοινωνικά δίκτυα τα οποία διαφέρουν μεταξύ τους. Θα ήταν καλό να δημιουργηθεί ένα κοινωνικό δίκτυο το οποίο να περιέχει όλα τα υπόλοιπα. Έτσι ένας χρήστης θα μπορεί να διαχειρίζεται τα προφίλ του εύκολα και αποδοτικά. Παράλληλα, θα βοηθήσει τους χρήστες να κατανοήσουν και να διαφυλάξουν τα προσωπικά τους στοιχεία έτσι ώστε να μην παρουσιάζουν πολλά δεδομένα προς τα έξω.

Έχοντας υπόψιν το κύριο πρόβλημα μας καθώς και μελλοντικές εργασίες που απορρέουν μέσα από τη λύση του πρόβλημα μας, προσπαθήσαμε να δημιουργήσουμε

[2]Πηγή: Growing social media, Pewinternet.org, LinkedIn.com

ένα εργαλείο που να μπορεί να κάνει ταύτιση των προφίλ ενός χρήστη με τα υπόλοιπα προφίλ του.

1.3 Στόχοι διπλωματικής και Δομή κειμένου

Σε αυτό το σημείο θα πρέπει να καθορίσουμε τους στόχους της διπλωματικής εργασίας. Πρώτιστος στόχος, είναι η εύρεση ενός αλγορίθμου που θα μπορεί με κάποιο ποσοστό επιτυχίας να αποφασίζει κατά πόσο δύο ή περισσότερα προφίλ χρηστών από διαφορετικά κοινωνικά δίκτυα ανήκουν στον ίδιο χρήστη ή όχι.

Στη συνέχεια αφού βρεθεί ο συγκεκριμένος αλγόριθμος, θα πρέπει να δημιουργηθεί ένα εργαλείο που να παρέχει αυτήν την δυνατότητα. Στην περίπτωσή μας, αυτό το εργαλείο θα είναι ένα web app όπου ο χρήστης θα βάζει ένα profile url και θα του επιστρέφονται τα προφίλ για όλα τα προφίλ που έχει ο χρήστης στα κοινωνικά δίκτυα με κάποιο ποσοστό επιτυχίας. Στη συνέχεια θα αναφέρουμε τα κεφάλαια του κειμένου και μία σύντομη περιγραφή για το κάθε ένα.

Κεφάλαιο 1

Σε αυτό το κεφάλαιο παρουσιάζονται διάφορα προβλήματα και το motivation που μας ώθησαν στο να προσπαθήσουμε να προτείνουμε μία λύση. Επίσης θα γίνει αναφορά σε κάποια από τα κοινωνικά δίκτυα και παρουσίαση κάποιων στατιστικών όσο αφορά την επιρροή των κοινωνικών δικτύων στις μέρες μας.

Κεφάλαιο 2

Στο κεφάλαιο 2 θα παρουσιαστούν δουλείες που έχουν γίνει και έχουν παρόμοια λογική. Αυτές οι εργασίες είτε προσπαθούν να λύσουν κάποιο άλλο πρόβλημα που έχει σχέση με τα κοινωνικά δίκτυα είτε προσπαθούν να λύσουν ένα παρεμφερή πρόβλημα.

Κεφάλαιο 3

Στο κεφάλαιο 3 θα γίνει αναφορά στο friendFeed το οποίο χρησιμοποιήθηκε για την συλλογή των απαραίτητων δεδομένων για την ανάλυση.

Κεφάλαιο 4

Στο κεφάλαιο 4 θα παρουσιαστεί η ανάλυση η οποία έχει γίνει και η προετοιμασία η οποία έγινε έτσι ώστε να έχουμε σωστά και δομημένα δεδομένα έτσι ώστε να μπορέσουμε να αναλύσουμε την συμπεριφορά των χρηστών.

Κεφάλαιο 5

Σε αυτό το κεφάλαιο θα παρουσιάσουμε τον αλγόριθμό μας με βάση την ανάλυση η οποία έχει γίνει.

Κεφάλαιο 6

Στο κεφάλαιο 6 θα γίνει παρουσίαση των αποτελεσμάτων που έχουμε βγάλει από τον αλγόριθμό που εισηγούμαστε και επίσης σύγκριση με ήδη υπάρχον μεθόδους.

Κεφάλαιο 7

Στο κεφάλαιο 7 θα γίνει παρουσίασή της διαδικτυακής εφαρμογής που θα προσφέρει αυτή την λειτουργιά τους χρήστες.

Κεφάλαιο 8

Και τέλος θα παρουσιαστεί η μελλοντική δουλεία και τα γενικά συμπεράσματα που έχουμε βγάλει.

Παραρτήματα

Έχουμε προσθέσει δύο παραρτήματα όπου στο πρώτο παράρτημα παρουσιάζεται ο κώδικά που υλοποιήθηκε στην R στα πλαίσια της ανάλυσης των δεδομένων μας και στο δεύτερο παράρτημα παρουσιάζονται τα εργαλεία που χρησιμοποιήθηκαν κατά την διάρκεια την εκπόνησης της διπλωματικής εργασίας.

1.4 Thesis' contribution

Μέσα από αυτή την διπλωματική εργασία προτείνουμε έναν αλγόριθμο, ο οποίος θα εξάγει όλες τις πληροφορίες για έναν συγκεκριμένο χρήστη από ένα κοινωνικό δίκτυο. Στη συνέχεια θα βρίσκει όλα τα πιθανά προφίλ που έχει αυτός ο χρήστης σε άλλα κοινωνικά δίκτυα και θα τα παρουσιάζει με μία πιθανότητα, η οποία θα εκφράζει την ακρίβεια της ταύτισης. Πιο συγκεκριμένα παρέχουμε:

1. Ταύτιση προφίλ κοινωνικών δικτύων
2. Ακριβής αποτελέσματα
3. Γρήγορη απόκριση αποτελεσμάτων
4. Υποσχόμενες προοπτικές για εργασίες που θα βασίζονται σε αυτή την εργασία

Κεφάλαιο 2

Related Work

2.1 Εισαγωγή

Ξεκινώντας αυτό το κεφάλαιο θα ήταν σωστό να αναφέρουμε το κύριο πρόβλημα που προσπαθούμε να λύσουμε. Το πρόβλημα μας, είναι αυτό της αναζήτησης κάποιου ατόμου στα κοινωνικά δίκτυα. Επίσης και η ώθηση που μας δόθηκε σκεπτόμενοι των μεγάλο αριθμό από μελλοντικές εργασίες που θα μπορούσαν να υλοποιηθούν έχοντας υπόψιν αυτή την εργασία, όπως αναπτύχθηκαν στο κεφάλαιο 1.

Συνεπώς, τώρα θα πρέπει να συνεχίσουμε σε μία έρευνα πάνω σε εργασίες που έχουν γίνει που έχουν στόχο την επίλυση παρόμοιων προβλημάτων με το πρόβλημα που προσπαθούμε να επιλύσουμε ή/και προβλήματα που έχουν άμεση σχέση με τα κοινωνικά δίκτυα.

Μέσα από αυτές τις εργασίες θα μπορέσουμε να δούμε τυχόν δυσκολίες που αντιμετώπισαν καθώς και τεχνικές που ίσως μας φανούν χρήσιμες στην πορεία. Πέραν αυτού θα μπορέσουμε να δούμε τυχόν βελτιώσεις που μπορεί να εισηγηθούμε και να τις εφαρμόσουμε στην δική μας προσέγγιση.

2.2 Categorization

Για καλύτερη παρουσίαση έχουμε κατηγοριοποιήσει τις εργασίες που έχουμε σε δύο κύριες κατηγορίες. Ο λόγος που προχωρήσαμε σε αυτό τον διαχωρισμό είναι για να έχουμε καλύτερη αντίληψη των εργασιών που ασχολήθηκαν καθαρά με το πρόβλημα του profile matching σε σχέση με τα υπόλοιπα που παρουσιάζουν τεχνικές εξόρυξης ή/και παρουσίασης πληροφοριών στα κοινωνικά δίκτυα.

Χαρακτηριστικά	Κοινωνικά Δίκτυα
Name	Both
Email	Both
Education	Facebook
employment	Facebook
Gender	Both
Age	MySpace
Country	Both
City	Both
Hometown	Facebook
Zip	Both

Πίνακας 2.1: Χαρακτηριστά χρηστών ανά Κοινωνικό Δίκτυο

2.2.1 Profile Matching

Αυτή η κατηγορία αναφέρεται σε άρθρα στα οποία επικεντρώθηκαν στο profile matching για διαφορετικά κοινωνικά δίκτυα. Παρουσιάζουν τεχνικές και αποτελέσματα των προσεγγίσεων τους.

Η εργασία των M.Motoyama G.Varghese [4] ανήκει σε αυτή την κατηγορία όπου συζητούν για ένα σύστημα αναζήτησης και ταύτισης χρηστών μεταξύ Facebook και MySpace.

Η γενική προσέγγιση που προτάθηκε ήταν να επιλέγουν ένα προφίλ από τα δύο αυτά κοινωνικά δίκτυα και να αναζητούν το αντίστοιχο προφίλ στο άλλο κοινωνικά δίκτυο, χρησιμοποιώντας τα χαρακτηριστικά των προφίλ. Έχουν αναφέρει ότι χρησιμοποιώντας την ηλεκτρονική διεύθυνση των χρηστών θα μπορούσαν να βγάλουν καλύτερα αποτελέσματα. Χρησιμοποίησαν την OCR engine που είναι μια τεχνική για αναγνώριση χαρακτήρων από εικόνα, για να μπορέσουν να πάρουν την ηλεκτρονική διεύθυνση από το Facebook. Τα χαρακτηριστικά και τα αποτελέσματα τις έρευνας τους παρουσιάζονται στου πίνακες 2.1 και 2.2 αντίστοιχα.

Οι διαφορές με την δικιά μας προσέγγιση είναι ότι εμείς χρησιμοποιούμε στα χαρακτηριστικά μας ένα είδος βάρους που αξιολογεί και ορίζει την σημαντικότητα του κάθε χαρακτηριστικού. Επίσης έχουμε προσθέσει την εικόνα προφίλ του χρήστη έτσι ώστε να έχουμε ακόμα ένα χαρακτηριστικό στα χέρια μας. Τα αποτελέσματα μας, όπου θα παρουσιαστούν σε μετέπειτα κεφάλαιο, είναι καλύτερα σε σχέση με το πιο πάνω άρθρο. Επίσης εμείς κάνουμε ανάλυση πάνω σε τέσσερα κοινωνικά δίκτυα σε σχέση με αυτή την δουλεία που επικεντρώνεται σε δύο.

Συνδυασμοί Χαρακτηριστικών	Ποσοστό επιτυχίας
Age,gender,location,name	1.7
Age , location, name	0.36
Country,gender,name	0.04
Country, name	6.96
Email	11.7
Gender,location,name	0.06
Href	5.7
Lacation,name	12.5
Name	37.82
Name,school	14.2
Name,work	3.6

Πίνακας 2.2: Συνδυασμός χαρακτηρισμών και ποσοστό επιτυχίας

Στη συνέχεια οι G.You κ.α [18] προσπαθούν να μελετήσουν αν μπορούν να ταυτίσουν το όνομα κάποιου χρήστη με τον λογαριασμό του στο Twitter. Και εισηγούνται ένα εργαλείο το οποίο να κάνει αυτή την λειτουργία.

Προτείνουν ένα εργαλείο που αποτελείτε από τρία βήματα:

1. Selecting candidate account
2. Ranking candidate
3. Computing Confidence

Για την επιλογή των υποψήφιων λογαριασμών παρουσιάζουν τρεις κατηγορίες:

E-type
: if the account name is the same with the query

A-type
: if the account name contains all the words of the query

S-type
: if the account name contains some of the words query

Η κάλυψη από αυτές τις κατηγορίες για το E-type 0,7933 για το A-type 0,9149 καιS-type με 0,9726

Για την κατάταξη των υποψηφίων χρησιμοποίησαν δύο μεθόδους:

Μέθοδος	F	Precision	Recall	Accuracy
S+SVM+NUL	0.6336	0.4839	0.9241	0.4839
S+SVM+REL	0.8062	0.7762	0.8403	0.8183
S+SVM+CNF	0.8148	0.7917	0.8415	0.8282
S+SVM+CNF+F1	0.8178	0.7856	0.8543	0.8299

Πίνακας 2.3: Αποτελέσματα SocialSearch: Enhancing Entity Search with Social Network Matching

1. Relevance

2. Popularity

Πήραν από τους λογαριασμούς στο Twitter, το όνομα, το id και τους σχετιζόμενους κόμβους και χρησιμοποίησαν A- E- S- type για να μετρήσουν το όνομα, το id και τη σχετικότητα.

Και στη συνέχεια για να μετρήσουν το πόσο διάσημοι είναι, μέτρησαν τον αριθμό των followers.Τα αποτελέσματα του πιο πάνω άρθρου παρουσιάζονται στον πίνακα 2.3. Η μέθοδός τους, κάνει χρήση του S-type σε συνδυασμό με το SVM (Support Vector Machine). Και έκαναν τις εξής προσπάθειες:

1. S-type+SVM+Null using no confidence scoring

2. S-type+SVM+Rel using relevance for confidence scoring

3. S-type+SVM+CNF using seperate confidence score for confidence scoring

4. S-type+SVM+CNF+F1 using CNF optimized for F1 measure

Οι διαφορές που έχουμε με αυτό το άρθρο είναι ότι χρησιμοποιούν μόνο το όνομα ενός χρήστη για να αξιολογήσουν κατά πόσο δύο χρήστες έχουν το ίδιο προφίλ. Το δικό μας μοντέλο παρουσιάζει μία προσέγγιση που κάνει χρήση μίας πλειάδας από χαρακτηριστικά για να εξασφαλίσει καλύτερα αποτελέσματα.

Επίσης κάνουν χρήση του SVM (Support Vector Machine) σε αντίθεση με εμάς που κάνουμε χρήση της logistic regression. Πιστεύουμε ότι έχουμε καλύτερα αποτελέσματα σε σχέση με αυτήν την εργασία και ο λόγος είναι ότι χρησιμοποιούμε logistic regression.

Τέλος, η εργασία των E. Raad R. Chbeir and A. Dipanda [1], όπου παρουσιάζουν ένα εργαλείο όπου δίνει την δυνατότητα στους χρήστες να βρίσκουν κατά πόσο δύο προφίλ ανήκουν στον ίδιο χρήστη.

Στα πλαίσια αυτής της δουλειάς επιλέχτηκαν το Facebook και το LinkedIn .

Προσπάθησαν να ερευνήσουν τρεις κύριες περιοχές:

1. Social network profile heterogeneity
2. Similarity measuring between attribute values
3. Decesion making about whether two profiles refer to the same person or not

Το framework τους αποτελείται από τέσσερα συστατικά:

1. FOAF Middleware
2. Similarity Functions Assignment
   (α′) Sysntactic-based similarity approaches
   (β′) Semantic-based similarity approaches
   (γ′) Senseless One-term attributes
   (δ′) Senseless Multi-terms attributes
3. Attribute Weight Assignment
4. Profile Matching

Το FOAF χρησιμοποιείται σαν ένας τρόπος αναπαράστασης των πληροφοριών των προφίλ από το framework . Κάθε χαρακτηριστικό έχει το δικό του βάρος όσο αφορά την σημαντικότητα από τα δύο κοινωνικά δίκτυα πχ. Στο LinkedIn το χαρακτηριστικό homepage είναι πιο σημαντικό από ότι στο Facebook.

Σε αυτό το framework χρησιμοποιούν DS function σαν την κύρια μέθοδο για να αποφανθεί κατά πόσο δύο προφίλ ανήκουν στον ίδιο χρήστη, αφήνοντας στον χρήστη την επιλογή να αλλάζει τις καθορισμένες ρυθμίσεις.

Αυτό το framework είναι ικανό να βρει κατα πόσον δύο προφίλ ανήκουν στον ίδιο χρήστη ή όχι. Επίσης χρησιμοποίησαν και άλλες τεχνικές για ταύτιση προφίλ όπως DS, BN, Avg, Min, and Max. Το καλύτερο αποτέλεσμα όσο αφορά το precision το είχε το DS. Δεύτερο ήταν το BN έπειτα το Avg και τέλος τα Min, and Max αντίστοιχα. Τα αποτελέσματα τις προσέγγισης τους παρουσιάζονται στον πίνακα 2.4. Οι διαφορές που έχουμε με αυτό το άρθρο είναι σχεδόν ίδιες με του προηγούμενου άρθρου όπου για να αξιολογήσουν κατά πόσο δύο χρήστες έχουν το ίδιο προφίλ χρησιμοποιούν DS function,σε αντίθεση με εμάς που κάνουμε χρήση της logistic regression.

Το δικό μας μοντέλο παρουσιάζει μία προσέγγιση που κάνει χρήση όλων των χαρακτηριστικών που μπορούμε να εξασφαλίσουμε από τα κοινωνικά δίκτυα για να εξασφαλίσουμε καλύτερα αποτελέσματα. Από την άλλη, αυτό το άρθρο χρησιμοποιεί μερικά χαρακτηριστικά και αφήνει στον χρήστη την επιλογή αυτών.

Μέθοδος	Precision	Recall
DS Function	0.96	0.70
Average	0.80	0.70
Naive Bayes	0.83	0.70
Min	0.80	0.77
Max	0.10	0.91

Πίνακας 2.4: Αποτελέσματα της εργασίας των E. Raad κ.α [1]

Επίσης όσο αφορά το βάρος που επιλέγεται, δεν καθορίζουν κάποιον συγκεκριμένο τρόπο εύρεσης αλλά αφήνουν τον χρήστη να αποφασίσει πόσο βάρος θα δώσει σε κάθε χαρακτηριστικό σε αντίθεση με την δική μας προσέγγιση που βρίσκει το βέλτιστο βάρος για κάθε χαρακτηριστικό.

2.2.2 General information

Σε αυτήν την ενότητα θα παρουσιάσουμε εργασίες που επικεντρώθηκαν σε τεχνικές εξόρυξης δεδομένων από τα κοινωνικά δίκτυα καθώς επίσης και επιπλέον τεχνικές που ασχολούνται με τον τρόπο παρουσιάσεις δεδομένων.

Η εργασία των L. Ding κ.α. [5], παρουσιάζουν τεχνικές παρουσίασης των πληροφοριών για να βελτιστοποιήσουν τόσο τον τρόπο παρουσίασης καθώς και την τρόπο αποθήκευσης των πληροφοριών.

Κάνουν έμφαση σε μία μεγάλη πλειάδα από οντολογίες στο διαδίκτυο. Παρουσιάζουν διαφορετικούς τύπους από οντολογίες όπως τα RDF, FOAF, DC, RDFS etc.

Παρουσιάζουν και τονίζουν τη σημαντικότητα του FOAF, το οποίο είναι το πιο δημοφιλή στο διαδίκτυο. Πολλές ιστοσελίδες, μπλοκ και κοινωνικά δίκτυα χρησιμοποιούν τέτοιου είδους αρχεία για να διατηρούν και να περιγράφουν μία οντότητα στο διαδίκτυο. (πχ.user, group, company etc). FOAF document είναι RDF/XML λεξιλόγιο για περιγραφή οντοτήτων.

Τέλος παρουσιάζουν τα κύρια χαρακτηριστικά που χρησιμοποιούνται για την παρουσίαση μίας οντότητας. Όπως φαίνονται πιο κάτω:

1. name
2. knows
3. homepage
4. mboxsh1sum

5. seeAlso
6. title
7. nickname
8. weblog
9. inbox
10. equivalentTo

Στην συνέχεια οι I. Polakis κ.α [7], παρουσιάζουν τρόπους προστασίας και ασφάλειας των προσωπικών δεδομένων και παράλληλα παρουσιάζουν τεχνικές για εξόρυξη διευθύνσεων ηλεκτρονικού ταχυδρομείου όπως η πιο κάτω:

Blind Harvesting

Πρώτιστος στόχος για το blind Harvesting είναι να συλλέγει όσες πιο πολλές διευθύνσεις ηλεκτρονικού ταχυδρομείου μπορεί με το να κάνει συνέχεια queries για ονόματα στο Google search engine.

Targeted Harvesting

Συλλέγει ονόματα από δημοφιλή fan/group pages από το Facebook και στη συνέχεια κάνει χρήση του blind Harvesting. Για περισσότερη ακρίβεια, παίρνει ζευγάρια από (nickname, names) από το Twitter και εξάγει emails με πρόθεμα αυτό που είναι ακριβός το ίδιο με το nickname και μετά να ψάχνει για αυτό στο Facebook με 43,4% precision

Επίσης γίνεται αναφορά σε τεχνικές που χρησιμοποιούνται από τους spammers:

1. Web crawling: From blogs, personal web pages but suffers from scalability and is very time consuming activity.
2. Crawling Archive Sites that contains emails
3. Malware
4. Malicious sites
5. Dictionary attacks

Άρθρο	Μεθοδολογία	Κοινωνικά Δίκτυα	Precision	Recall
M.Motoyama G.Varghese [4]	Attribute based	Facebook/MySpace	37.82%	-
G.You κ.α [18]	Name matching+SVM	Twitter	79%	85%
E.Raad R.Chbeir and A.Dipanda [1]	Attribute based+DS function	Facebook/LinkedIn	96%	70%
Our Approach	Attribute based+Logistic Regression	Facebook/Flickr	58%	55%
Our Approach	Attribute based+Logistic Regression	Twitter/LinkedIn	85%	79%

Πίνακας 2.5: Κύρια χαρακτηριστικά εργασιών που ασχολήθηκαν με profile matching σε σχέση με την δική μας προσέγγιση

2.3 Συμπεράσματα

Κάθε άρθρο έχει τα δικά του χαρακτηριστικά και τις δικές του προσεγγίσεις. Παρουσιάζουμε, στους πίνακες αντίστοιχα, τα κυρία χαρακτηριστικά τόσο για τα άρθρα [1] που ασχολούνται με το profile matching, όπως φαίνονται στον πίνακα 2.5, και όσο και για τα άρθρα που μας παρέχουν τεχνικές για εξαγωγή και παρουσίαση πληροφοριών, πίνακα 2.6.

Άρθρο	Είδος	Κοινωνικά Δίκτυα
L. Ding κ.α. [5]	Παρουσίαση Πληροφοριών FOAF	Γενικά
I. Polakis κ.α [7]	Security and email harvesting	Twitter/Facebook

Πίνακας 2.6: Κύρια χαρακτηριστικά εργασιών που ασχολήθηκαν με τεχνικές εξόρυξης ή/και παρουσίασης πληροφοριών

Μέσα από τα πιο πάνω άρθρα [1] [18] [4] [7] [5], παρουσιάζονται κάποια προβλήματα.

[1] Μελλοντική Εργασία η υλοποίηση τον εργασιών αυτών

Παρατηρήσαμε ότι τα πιο πάνω άρθρα επικεντρώνονται στο όνομα και στην διεύθυνση του ηλεκτρονικού ταχυδρομείου του χρήστη για ταύτιση, [18] [18] με αποτέλεσμα να μην έχουν καλά αποτελέσματα λόγο του ότι οι χρήστες δεν χρησιμοποιούν τον ίδιο λογαριασμό ηλεκτρονικού ταχυδρομείου σε όλα τα κοινωνικά δίκτυα που έχουν λογαριασμό. Αυτό θα προσπαθήσουμε να το επιλύσουμε εισάγοντας και άλλα χαρακτηριστικά έτσι ώστε να έχουμε μια ολοκληρωμένη και σφαιρική παρουσίαση ενός προφίλ. Συνεπώς οι τεχνικές που αναφέρθηκαν στο άρθρο [7] δεν θα τις χρησιμοποιήσουμε.

Όσο αφορά τα χαρακτηριστικά που εισηγούνται στα άρθρα [1] [18], θα προσθέσουμε σε αυτά, την εικόνα προφίλ και ότι χαρακτηριστικό είναι προσβάσιμο και διαθέσιμο. Επίσης δεν παρουσιάζεται κάποια μέθοδος που να αξιολογεί και να βρίσκει την σημαντικότητα των χαρακτηριστικών που συμβάλουν στην ταύτιση των προφίλ. Αυτό θα προσπαθήσουμε να το λύσουμε με την χρήση της logistic regression που θα υπολογίσει την σημαντικότητα και το βέλτιστο βάρος κάθε χαρακτηριστικού.

Κεφάλαιο 3

FriendFeed

3.1 Εισαγωγή

Αμέσως μετά την έρευνα πάνω σε σχετικά άρθρα καθώς και στα προβλήματα που παρατηρήσαμε πάνω σε αυτά, θα πρέπει να βρούμε ένα εργαλείο που να μας επιτρέπει να εξάγουμε δεδομένα για να κατασκευάσουμε το ground truth που θα περιέχει συνδυασμούς προφίλ, όπου θα γνωρίζουμε από την αρχή ότι ανήκουν στον ίδιο χρήστη. Αυτό το εργαλείο είναι το FriendFeed που αποτελεί και το κύριο εργαλείο εξαγωγής δεδομένων.

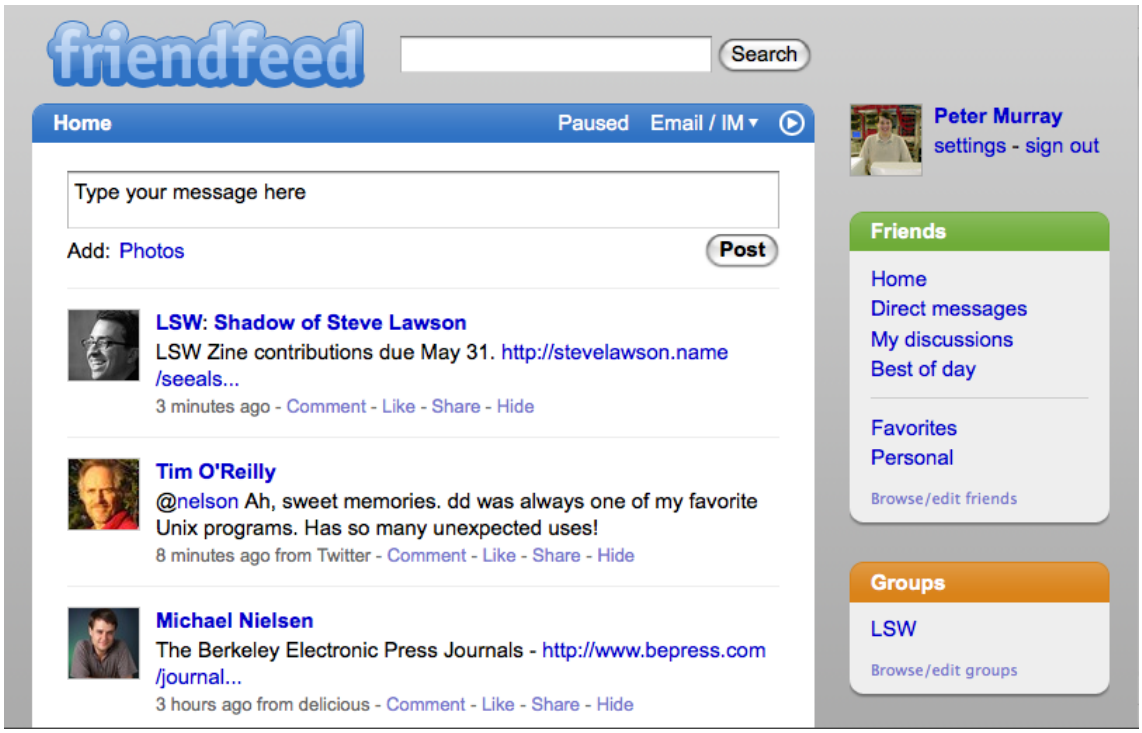
3.2 Aggregate Tools

Aggregate Tools είναι τα εργαλεία τα οποία δίνουν την δυνατότητα στους χρήστες να ενώνουν τα κοινωνικά τους δίκτυα σε ένα λογαριασμό, έτσι ώστε κάνοντας share ή tweet ή οποιαδήποτε δημοσίευση, να γίνεται αυτόματα και σε πραγματικό χρόνο δημοσίευση στον λογαριασμό του χρήστη σε ένα τέτοιο εργαλείο.

Το FriendFeed είναι ένα από τα εργαλεία τα οποία χρησιμοποιήσαμε για να κάνουμε αυτήν την εργασία. Στο σχήμα 3.1 παρουσιάζονται τα feeds που κάνουν οι χρήστες καθώς και από πιο κοινωνικό δίκτυο έγινε.

Σχήμα 3.2: FriendFeed logo

Το FriendFeed είναι ένα εργαλείο το οποίο ενώνει σε πραγματικό χρόνο τα τρέχων feeds των χρηστών από τα κοινωνικά δίκτυα, μπλοκ, μικρο-μπλοκ καθώς επίσης και από διάφορα RSS/Atom. Χρησιμοποιόταν για να δημιουργήσει μία ροή από πληροφορία καθώς και για να δημιουργεί νέες συζητήσεις και σχόλια για διάφορες δημοσιεύσεις.



Σχήμα 3.1: Users' profile in FriendFeed

Στόχος του FriendFeed ήταν να μετατρέψει τα δεδομένα που υπάρχουν στο διαδίκτυο, σε χρήσιμη πληροφορία και εύκολα διαχειρίσιμη από όλους. Η υπηρεσία αυτή έχει τερματιστεί στις 9-Απρ-2015 αφού εξαγοράστηκε από το Facebook [13]

Η επιλογή του FriendFeed έγινε μετά από σκέψη και αναζήτηση πάνω σε κοινωνικά δίκτυα που συναθροίζουν τα προφίλ των χρηστών. Ήταν ένα από τα πιο δημοφιλή κοινωνικά δίκτυα αυτής της κατηγορίας που παρείχε και το απαιτούμενο API, δηλαδή μεθόδους για να μπορείς να εξάγεις πληροφορία.

3.3 Μεθοδολογία

Για να μπορέσουμε να αξιοποιήσουμε αυτήν την πληροφορία χρησιμοποιήσαμε το API του FriendFeed το οποίο μας παρείχε μεθόδους για να μπορέσουμε να εκμεταλλευτούμε και να αξιοποιήσουμε αυτή την πηγή πληροφοριών.

Αρχικά χρησιμοποιήσαμε το public stream για να συλλέξουμε δημοσιεύσεις των χρηστών. Το public stream περιείχε όλες τις πρόσφατες δημοσιεύσεις των χρηστών σε μορφή αντικειμένου json. Κάθε αντικείμενο του public stream ήταν ένα feed όπου περιείχε πληροφορίες σχετικά με την δημοσίευση.

Συνολικά έχουμε μαζέψει περισσότερες από ένα εκατομμύριο δημοσιεύσεις χρηστών χρησιμοποιώντας http requests. Αρχικά για κάθε ένα από τα feeds μπορούσαμε να

εξάγουμε τις ακόλουθες πληροφορίες.

1. Username
2. Body of the post
3. Comments
4. Likes
5. Geo location

Στη συνέχεια για να εξορύξουμε τα κοινωνικά δίκτυα τα οποία έχει ένας χρήστης χρησιμοποιήσαμε ιχνηλάτες [17] για να μπορέσουμε να εξάγουμε αυτή την πληροφορία. Ιχνηλάτες είναι προγράμματα τα οποία ιχνηλατούν τον ιστό με κάποιο συγκεκριμένο σκοπό. Στη δική μας περίπτωση, ήταν να εξάγει τα κοινωνικά δίκτυα τα οποία έχει συναθροίσει ένας χρήστης.

Η χρήση ιχνηλατών είναι σχετικά μία χρονοβόρα διαδικασία με αποτέλεσμα να καθυστερήσουμε να πάρουμε αυτή την πληροφορία. Η λύση η οποία υλοποιήσαμε όπως φαίνεται και στο σχήμα 3.3 ήταν να δημιουργήσουμε ένα multithread enviroment για να μειώσουμε το χρόνο περάτωσης της εργασίας.

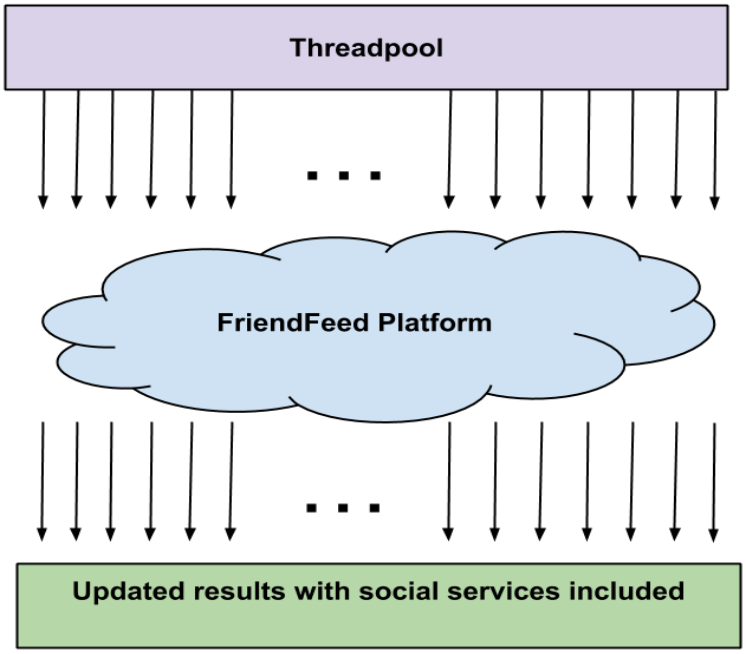
Κάθε ένα από τα threads που ξεκινούσε περιείχε σαν είσοδο ένα user profile url όπου έμπαινε στο προφίλ του χρήστη και παίρνοντας όλο το html source αναζητούσε τα κοινωνικά δίκτυα που έχει συναθροίσει ένας χρήστης.

Στο τέλος κάθε thread επέστρεφε τα κοινωνικά δίκτυα που έχει κάποιος χρήστης. Με αυτόν τον τρόπο είχαμε καταφέρει να δημιουργήσουμε μια συλλογή από δεδομένα που να περιείχαν πληροφορίες για τους χρήστες του FriendFeed και επιπλέον τα κοινωνικά δίκτυα τα οποία έχουν συναθροίσει.

3.3.1 Ανάλυση δεδομένων FriendFeed

Αφού τώρα έχουμε τα κοινωνικά δίκτυα για κάθε ένα από τους χρήστες θα μπορέσουμε να αναλύσουμε και να βρούμε τα κοινωνικά δίκτυα που θα χρησιμοποιήσουμε για να δημιουργήσουμε τον αλγόριθμό μας. Μέσα από αυτά τα δεδομένα που έχουμε συλλέξει, έχουμε τα ακόλουθα κοινωνικά δίκτυα.

1. Facebook
2. Twitter
3. LinkedIn



Σχήμα 3.3: Multithread enviroment to retrive social services

4. Flickr
5. Tumblr
6. Instagram

3.3.2 Επιλογή Δικτύων

Στον πίνακα 3.3 παρουσιάζονται ο αριθμός των χρηστών για κάθε ένα από τα κοινωνικά δίκτυα που έχουν οι χρήστες. Για το Facebook έχουμε 4778 χρήστες, για το Twitter έχουμε 10233, 3684 χρήστες για το LinkedIn, για το Flickr έχουμε 7131, για το Tumblr 560 και τέλος για το Instagram 1014.

Στη συνέχεια πρέπει να επιλέξουμε τους συνδυασμούς των κοινωνικών δικτύων που θα χρησιμοποιήσουμε για να εφαρμόσουμε τον αλγόριθμο μας. Επιλέξαμε να συνδυάσουμε το Facebook με το Flickr και στη συνέχεια το Twitter με το LinkedIn, ο λόγος που επιλέχτηκαν αυτοί οι συνδυασμοί είναι επειδή είχαμε τους περισσότερους χρήστες. Στον πίνακα 3.4 παρουσιάζονται οι χρήστες που έχουν συνδυασμούς από κοινωνικά δίκτυα.

3.3.3 Useful Facts for Flickr

Στη συνέχεια αναλύσαμε την συμπεριφορά των χρηστών στο Flickr που έχουμε συλλέξει, με πολύ ενδιαφέροντα αποτελέσματα. Η πόλη με τους περισσότερους χρήστες με βάση τα δεδομένα τα οποία έχουμε συλλέξει είναι η Αμερική με 734 χρήστες τα αποτελέσματα παρουσιάζονται στον πίνακα 3.1. Παρατηρούμε ότι το 46% των χρηστών που έχουν Flickr δεν παρέχουν την τοποθεσία τους.

Στη συνέχεια βρήκαμε τα πιο συνηθισμένα ονόματα για να δούμε αν θα επηρεάσουν τα τελικά μας αποτελέσματα. Στα δεδομένα μας βρήκαμε ότι το πιο συνηθισμένο επίθετο είναι το Smith, ενώ όσο αφορά το όνομα είναι τα David, Chris, Andrew, John, Alex.

Flickr-FriendFeed Coverage

Έπειτα προσπαθήσαμε να ανακαλύψουμε εάν υπάρχει κάποια σύνδεση μεταξύ των followings in friendfeed με τους followings in flickr. Μέσα από την ανάλυση του γράφου των κοινωνικών δικτύων καταλήξαμε ότι κάθε χρήστης στο Friendfeed που έχει Flickr profile έχει κατά μέσο όρο 14,39 χρήστες όπου τους ακολουθεί και στα δύο κοινωνικά δίκτυα. Για να καταλήξουμε στο πιο πάνω αποτέλεσμα ακολουθήσαμε μια συγκεκριμένη διαδικασία όπως φαίνεται στον ψευδοκώδικα Algorithm 1.

Algorithm 1 Coverage Flickr FriendFeed

```
1:  procedure CAVARAGE
2:      N ← Data set with Flickr users
3:      W ← Data set with FriendFeed users
4:      coutner ← 0
5:      coverage ← 0
6:      for each entry i ∈ W do
7:          Take the set of following, S
8:          for each following j ∈ S do
9:              if j ∈ N then
10:                 counter ← counter + 1
11:         coverage ← coverage + (counter/sizeOf(S))
12:     return coverage
```

3.3.4 Useful Facts for Facebook

Στη συνέχεια αναλύσαμε τους χρήστες που είχαμε για το Facebook, για να μπορέσουμε να αναγνωρίσουμε κάποια συμπεριφορά που θα μας φαινόταν χρήσιμη για το μέλλον.

Χώρα	Χρήστες
Αμερική	734
Ισπανία	222
Ιαπωνία	122
Ηνωμένο Βασίλειο	109
Ιταλία	77
Γερμανία	57
Καναδάς	43
Γαλλία	40
Άλλο	1312
N/A	3870

Πίνακας 3.1: Χρήστες ανά χώρα για Flickr

Πληροφορίες	Χρήστες
Χωρίς Όνομα	55
Χωρίς φύλο	126
Χωρίς τοποθεσία	55
Χωρίς περιγραφή	4189
Άνδρες	3225
Γυναίκες	893

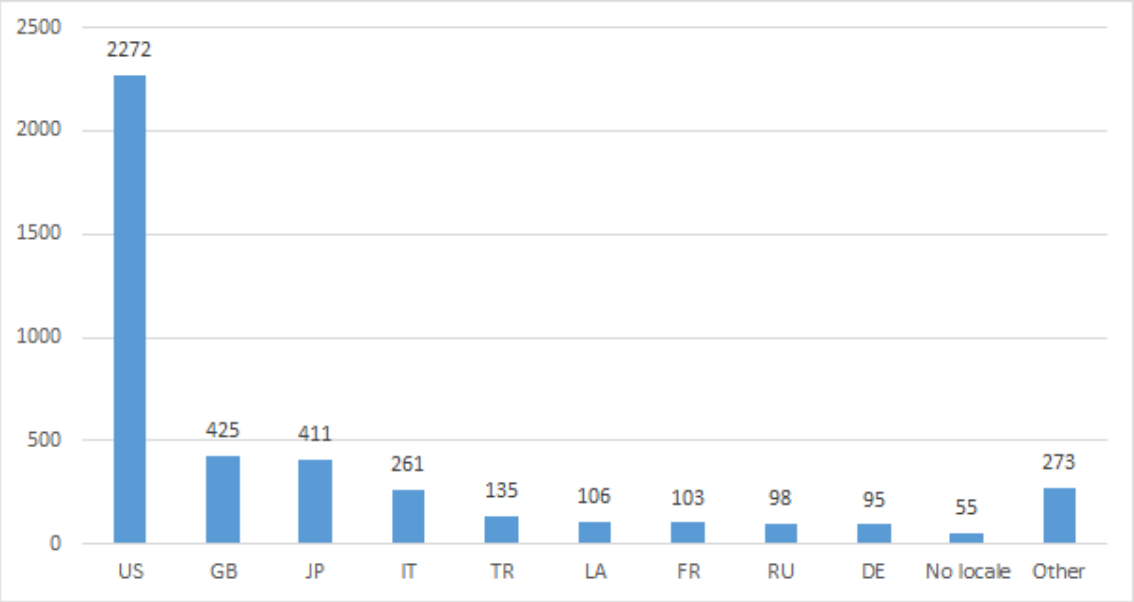
Πίνακας 3.2: Πληροφορίες από ανάλυση των δεδομένων του Facebook

Έχουμε βρει την χώρα με τους περισσότερους χρήστες(βλέπε Σχ. 3.4), η οποία είναι η Αμερική με 2272 χρήστες, σύμφωνα με τα δεδομένα τα οποία κατέχουμε. Στη συνέχεια παρουσιάζουμε κάποια γενικά στατιστικά όπως φαίνεται στον πίνακα 3.2.

Επίσης έχουμε βρει τα πιο συνηθισμένα επίθετα, τα οποία είναι Smith, Brown,Johnson και τα πιο συνηθισμένα ονόματα τα οποία είναι τα David, Chris ,John, Michael

3.4 Ground truth

Μετά τον καθορισμό των συνδυασμών των κοινωνικών δικτύων, πρέπει να συλλέξουμε τα χαρακτηριστικά των χρηστών από τα προφίλ που έχουν στα επιλεχθέντα κοινωνικά δίκτυα. Έτσι ώστε να βρεθούν τα χαρακτηριστικά τα οποία θα χρησιμοποιηθούν στον αλγόριθμο. Το ground truth το οποίο συλλέξαμε χρησιμοποιώντας το friend feed



Σχήμα 3.4: Facebook Users per Country

Κοινωνικά Δίκτυα	Αριθμός Χρηστών
Facebook	4778
Twitter	10233
LinkedIn	3684
Flickr	7131
Tumblr	560
Instagram	1014

Πίνακας 3.3: Αριθμός χρηστών ανά Κοινωνικό Δίκτυο

Συνδυασμός Δικτύων	Αριθμός Χρηστών
Facebook - Flickr	2292
Twitter - LinkedIn	3587
Twitter - Facebook	2113
Tumblr - Facebook	132
Instagram - Tumblr	203
Flickr - Twitter	2159

Πίνακας 3.4: Αριθμός χρηστών ανά συνδυασμό Κοινωνικών Δικτύων

aggregator, αποτελείται από χρήστες που έχουν λογαριασμό σε δύο ή/και περισσότερα, από τα ακόλουθα κοινωνικά δίκτυα.

1. Facebook
2. Twitter
3. LinkedIn
4. Flickr

3.4.1 Facebook - Flickr

Όπως φαίνεται και από τον πίνακα 3.4 έχουμε σύνολο 2292 χρήστες που έχουν λογαριασμό στο Facebook και Flickr. Συνεπώς γνωρίζουμε από την αρχή πια προφίλ έχουν τον ίδιο κάτοχο, αυτό που δεν γνωρίζουμε είναι πια προφίλ δεν έχουν τον ίδιο κάτοχο.

Αρχικά, προσηλωθήκαμε στο συνδυασμό Facebook και Flickr. Συλλέξαμε ότι πληροφορία μπορούσαμε να βρούμε για αυτούς τους χρήστες χρησιμοποιώντας τα API των κοινωνικών δικτύων.

Ολοκληρώνοντας την συλλογή των πληροφοριών για όλους τους χρήστες που έχουν Facebook και Flickr καταφέραμε να χτίσουμε ένα dataset το οποίο περιέχει για έναν χρήστη όλες τις πληροφορίες και για τα δύο προφίλ του.

Έχοντας αυτές πληροφορίες, προσπαθήσαμε να επεκτείνουμε τα δεδομένα μας με το να εισάγουμε προφίλ τα οποία δεν ανήκουν στον ίδιο χρήστη χρησιμοποιώντας το API του Facebook και το πραγματικό όνομα του χρήστη από το Flickr.

Αναλυτικά, κάναμε αιτήσεις στο graph API του Facebook όπου ζητούσαμε να μας επιστρέψει τα πρώτα 100, αν υπάρχουν, προφίλ που έχουν το ίδιο όνομα με αυτό του χρήστη στο Flickr.

Αν μέσα σε αυτά τα <=100 προφίλ περιέχεται το προφίλ το οποίο γνωρίζουμε από τα αρχικά μας δεδομένα ότι ανήκει στον ίδιο χρήστη τότε αναγράφουμε ότι έχουμε hit και το αναπαριστάμε με 1 αλλιώς βάζουμε 0.

Η μορφή που έχει το ground truth, μετά από την εισαγωγή των επιπλέον, λανθασμένων προφίλ που δεν ανήκουν στον ίδιο χρήστη παρουσιάζεται στον πίνακα 3.5

HIT/NO HIT	Προφίλ Facebook	Προφίλ Flickr
1	Χρήστης Α	Χρήστης Α
0	Χρήστης Α	Χρήστης Β
0	Χρήστης Α	Χρήστης Γ

Πίνακας 3.5: Μορφή του Ground Truth για Facebook - Προφίλ Flickr

3.4.2 Twitter - LinkedIn

Με την ίδια λογική, αρχικά θα βρούμε τους χρήστες που έχουν Twitter και LinkedIn χρησιμοποιώντας τα αρχικά δεδομένα τα οποία έχουμε συλλέξει από το FrienFeed.

Αφού συλλέξουμε τους χρήστες θα πάρουμε τα προφίλ τους και χρησιμοποιώντας τα API των κοινωνικών δικτύων θα εξάγουμε ότι πληροφορία είναι προσβάσιμη.

Συνεπώς έχουμε τα χαρακτηριστικά του ίδιου χρήστη και για τα δύο προφίλ που έχει. Στη συνέχεια θα πρέπει να γίνει επέκταση των δεδομένων εισάγοντας προφίλ τα οποία δεν ανήκουν στον ίδιο χρήστη. Αυτή τη φορά χρησιμοποιήσαμε το API του Twitter και με βάση το όνομα από το προφίλ στο LinkedIn, κάναμε ερωτήματα με βάση το όνομα και μας επέστρεφε 1000 προφίλ, όμως αναγκαστήκαμε να περιορίσουμε αυτό τον αριθμό στα 100.

Αν μέσα σε αυτά τα <=100 προφίλ περιέχεται το προφίλ το οποίο γνωρίζουμε από τα αρχικά μας δεδομένα ότι ανήκει στον ίδιο χρήστη τότε αναγράφουμε ότι έχουμε hit και το αναπαριστάμε με 1 διαφορετικά βάζουμε 0. Η μορφή που έχει το ground truth, μετά από την εισαγωγή των επιπλέον, λανθασμένων προφίλ που δεν ανήκουν στον χρήστη παρουσιάζεται στον πίνακα 3.6

HIT/NO HIT	Προφίλ Twitter	Προφίλ LinkedIn
1	Χρήστης Α	Χρήστης Α
0	Χρήστης Α	Χρήστης Β
0	Χρήστης Α	Χρήστης Γ

Πίνακας 3.6: Μορφή του Ground Truth για Twitter - Προφίλ LinkedIn

Κεφάλαιο 4

Ανάλυση Δεδομένων

4.1 Εισαγωγή

Στη συνέχεια ακολούθησε ανάλυση στα δεδομένα τα οποία έχουμε συλλέξει από το FriendFeed σε συνδυασμό με τα δεδομένα που έχουμε συλλέξει και από τα τέσσερα κοινωνικά δίκτυα που επικεντρωθήκαμε. Σε αυτό το κεφάλαιο θα παρουσιαστεί η ανάλυση που έγινε και τα χρήσιμα συμπεράσματα που εξήχθησαν από αυτήν. Καθώς και δυσκολίες οι οποίες προέκυψαν και με πιο τρόπο επιλύθηκαν

4.2 Ανάλυση Facebook - Flickr

Σε αυτή την ενότητα θα παρουσιάσουμε την ανάλυση και τα αποτελέσματα της ανάλυσης πάνω στα δεδομένα τα οποία έχουμε για τους χρήστες που έχουν Facebook και Flickr.

4.2.1 Χαρακτηριστικά Facebook - Flickr

Αρχικά πρέπει να γίνει επιλογή των χαρακτηριστικών που θα επιλεγούν από τα δεδομένα τα οποία έχουμε για να γίνει η κατηγοριοποίηση. Τα χαρακτηριστικά τα οποία έχουν χρησιμοποιηθεί, ήταν αρχικά το username, η τοποθεσία, το όνομα και η εικόνα προφίλ του χρήστη.

Ο λόγος στον οποίο έχουμε περιορισμένο αριθμό από μεταβλητές είναι η αυστηρή πολιτική που ακολουθεί το Facebook όσο αφορά το privacy των χρηστών. Στη συνέχεια λόγω αναβάθμισης του Graph Explorer API από την έκδοση 2.0 στην 2.1, έχουμε αφαιρέσει από την λίστα επιλογών το username του χρήστη διότι έχει απαγορευτεί στο username του χρήστη να γίνεται διαθέσιμο.

Χαρακτηριστικό	Είδος	Αλγόριθμοι
Όνομα Χρήστη	Συμβολοσειρά	Ngram/Levenshtein/Jaro
Τοποθεσία Χρήστη	Συμβολοσειρά	Ngram/Levenshtein/Jaro
Εικόνα Προφίλ Χρήστη	Εικόνα	Pixel Wise/Histogram/Data Correlation

Πίνακας 4.1: Δομή, Είδος, Πιθανοί Αλγόριθμοι για κάθε χαρακτηριστικό

Στη συνέχεια αφού απομονώσαμε τα στοιχεία τα οποία θα συμβάλουν στην δημιουργία και υλοποίηση του αλγορίθμου μεταξύ του Facebook και του Flickr, έπρεπε να αναλύσουμε τη δομή αυτόν των χαρακτηριστικών. Στον πίνακα 4.1 παρουσιάζονται η δομή, το είδος του κάθε χαρακτηριστικού καθώς και οι πιθανοί αλγόριθμοι για μέτρηση της αποστάσεις.

Όσο αφορά τους αλγορίθμους που προτείνονται:

Ngram

Ο Ngram χρησιμοποιείται για να κάνει ταύτιση πάνω σε συμβολοσειρες. Μετατρέπει μια σειρά από σύμβολα σε πιο μικρά σύνολα από n-gram. Με αυτό τον τρόπο επιτρέπει πιο αποδοτικό τρόπο σύγκρισης. Οι William B. Cavnar John M. Trenkle [8], εξηγούν και αναλύουν τον τρόπο που ο n-gram δουλεύει.

Levenshtein

Η Levenshtein [10] απόσταση είναι μία μετρική για συμβολοσειρές για να μετρήσει της διαφοράς μεταξύ τους. Ορίζεται σαν ο ελάχιστος αριθμός από τροποποιήσεις ενός χαρακτήρα που χρειάζεται να γίνει σε μια συμβολοσειρα, είτε διαγραφής είτε πρόσθεσης είτε αντικατάστασης,για να μετατραπεί μία λέξη στην άλλη. Πίο αναλυτικά παρουσιάζεται στην Βικιπεδια .

Jaro

Ο Jaro [14] αποτελεί την καλύτερη επιλογή για μικρού μήκους συμβολοσειρές. Η απόσταση Jaro ορίζεται ως εξής:

$$d_j = \frac{1}{3}(\frac{m}{|s1|} + \frac{m}{|s2|}\frac{m-t}{m}) \tag{4.1}$$

Όπου:

m

Ο αριθμός των χαρακτήρων που ταιριάζουν μεταξύ των συμβολοσειρών

t

ο αριθμός των μετατοπίσεων των χαρακτήρων

s1 , s2

Οι συμβολοσειρές για να υπολογίσουμε την απόσταση μεταξύ τους

Pixel Wise

Οι προσέγγιση του pixel wise είναι να ελέγχουμε μία εικόνα pixel per pixel. Δηλαδή, ελέγχουμε pixel προς pixel τις εικόνες που έχουμε και σημειώνουμε πόσα pixel είναι τα ίδια, προχωράμε μέχρι να ελέγξουμε όλα τα pixels. Για να δούμε πόσο ίδιες είναι δύο εικόνες, διαιρούμε των αριθμό των pixel που είναι ίδια ως προς των αριθμό όλων των pixels.

Histogram

Το histogram μίας εικόνας είναι μία γραφική όπου όπως φαίνεται και στο σχήμα 4.1, παρουσιάζονται, η συχνότητα των pixel με συγκεκριμένη ένταση. Το histogram είναι η αναπαράσταση μίας εικόνας. Συνεπώς η προσέγγιση του histogram matching [9] είναι να βρίσκει κατά πόσο δύο histogram είναι ίδια, πιο αναλυτική περιγραφή παρουσιάζεται στην Βικιπαίδεια.

Data Correlation

Το Digital Image Correlation (DIC) [12] βασίζεται στη μεγιστοποίηση του συντελεστή συσχέτισης που προσδιορίζεται με την εξέταση της έντασης των pixel σε δύο ή περισσότερες αντίστοιχες εικόνες και εξάγουν την χαρτογράφησης της παραμόρφωσης που σχετίζεται με τις εικόνες.
Η συσχέτιση εικόνας ορίζεται ως εξής:

$$r_ij(u, v, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial v}{\partial x}, \frac{\partial v}{\partial y}) = 1 - \frac{\sum_i \sum_j [F(x_i, y_j) - \bar{F}] - [G(x_i*, y_j*) - \bar{G}}{\sqrt{\sum_i \sum_j [F(x_i, y_j) - \bar{F}]^2 - [G(x_i*, y_j*) - \bar{G}]^2}} \tag{4.2}$$

Όπου:

F(xi,yj)

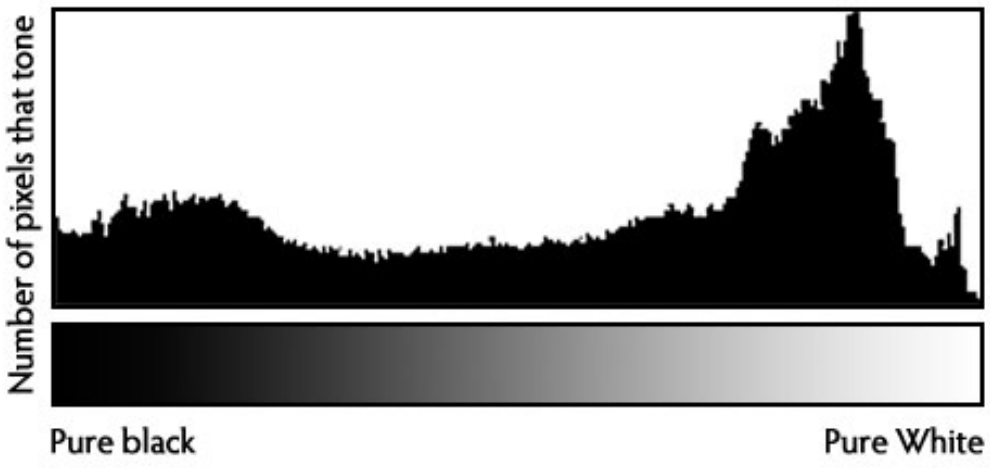
Η ένταση του πιξελ στη συγκεκριμένη θέση

G(xi*,yj*)

Η ένταση του πιξελ στη συγκεκριμένη θέση

$\bar{F}$, $\bar{G}$

Η μέση τιμή για τος πίνακες F G αντίστοιχα.



Σχήμα 4.1: Example of an image histogram

4.2.2 Users' Stats

Μέσα από τα δεδομένα που επιλέχτηκαν για την υλοποίηση του αλγορίθμου απορρέουν κάποιες σημαντικές παρατηρήσεις. Το σύνολο των εγγραφών που έχουμε είναι 12495, προσπαθήσαμε από αυτό το σύνολο να δούμε αν υπάρχει κάποια συμπεριφορά που να είναι σημαντική.

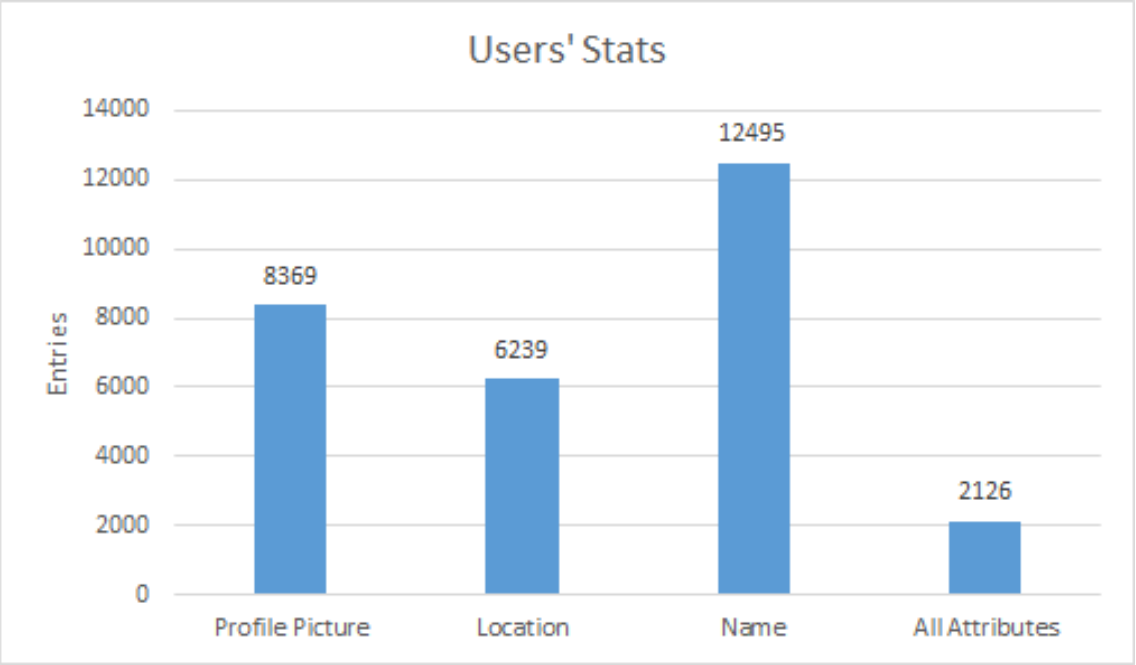
Παρατηρήσαμε ότι από τις 12495 εγγραφές οι 8369 έχουν εικόνα προφίλ, οι 6239 έχουν τοποθεσία χρήστη, και οι 2126 εγγραφές έχουν και όλα τα χαρακτηριστικά που όπως φαίνεται και στο σχήμα 4.2. Ως εκ τούτου, ο αλγόριθμός μας θα εφαρμοστεί πάνω στις 2126 εγγραφές.

4.2.3 Επιλογή Αλγορίθμων από Ανάλυση

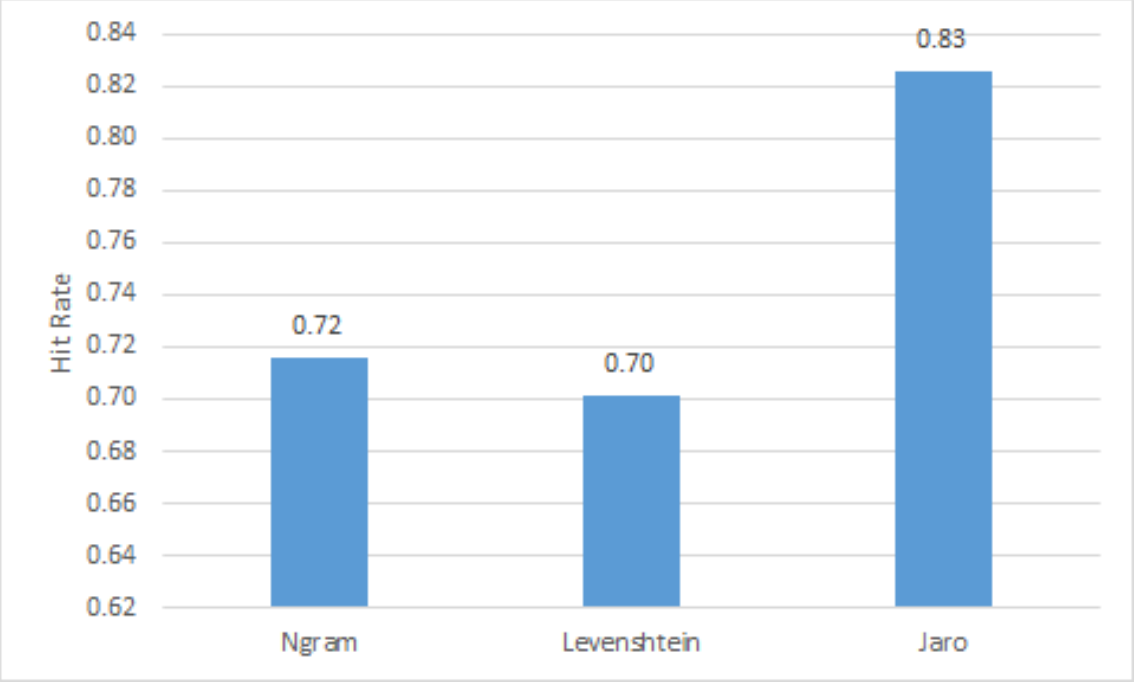
Μετά από τις παρατηρήσεις που κάναμε πάνω στα δεδομένα μας και στην επιλογή των χαρακτηριστικών που έγινε θα πρέπει να βρούμε τους τελικούς αλγορίθμους που θα χρησιμοποιήσουμε στον αλγόριθμο για την εύρεση των αποστάσεων/ομοιοτήτων μεταξύ των χαρακτηριστικών.

Για να καταλήξουμε σε μια απόφαση πήραμε τυχαίο δείγμα 700 χρηστών και εφαρμόσαμε πάνω σε αυτό τους αλγόριθμους για συμβολοσειρες και τους αλγορίθμους για εικόνες. Τα αποτελέσματα φαίνονται στα σχήματα 4.3 και 4.4 αντίστοιχα.

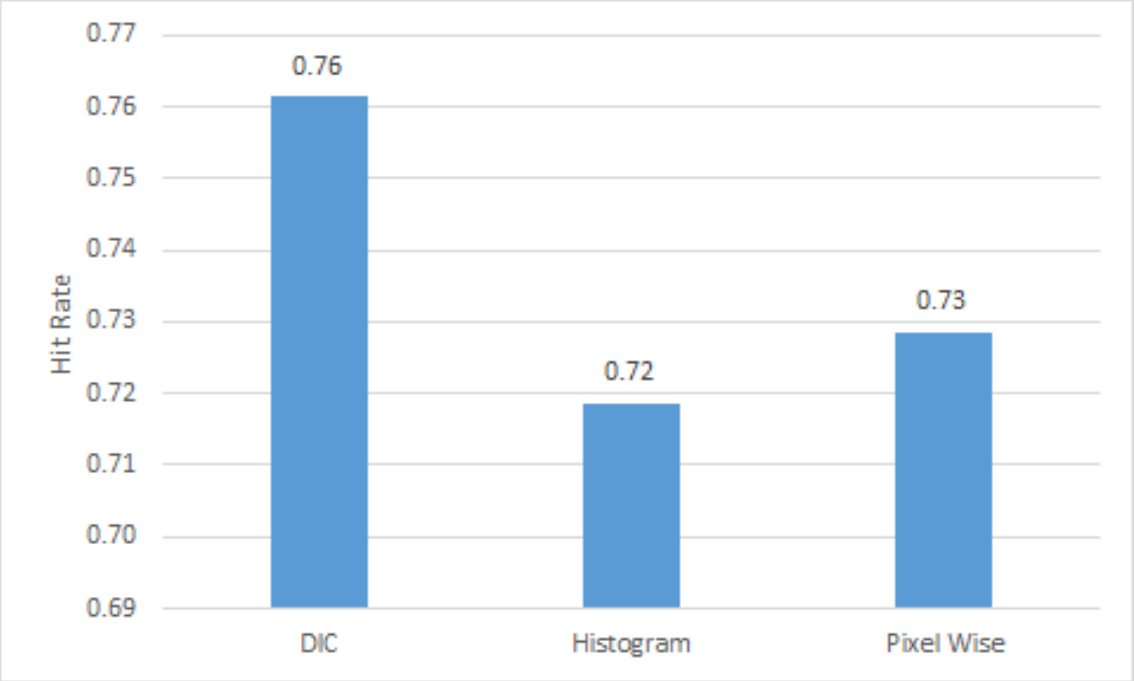
Έχοντας υπόψιν τα αποτελέσματα των γραφημάτων στα σχήματα 4.3 και 4.4 για τους string base algorithms και image base algorithms αντίστοιχα, οι προτεινόμενοι



Σχήμα 4.2: Users' Stats



Σχήμα 4.3: Hit rate of string base algorithm for Facebook and Flickr



Σχήμα 4.4: Hit rate of image base algorithm for Facebook and Flickr

Χαρακτηριστικό	Είδος	Αλγόριθμοι
Όνομα Χρήστη	Συμβολοσειρά	Jaro
Τοποθεσία Χρήστη	Συμβολοσειρά	Jaro
Εικόνα Προφίλ Χρήστη	Εικόνα	Data Correlation

Πίνακας 4.2: Προτεινόμενοι Αλγόριθμοι

αλγόριθμοι παρουσιάζονται στον πίνακα 4.2

4.2.4 Resampling

Λόγω του ότι στο ground truth μας τα προφίλ τα οποία γνωρίζουμε ότι δεν είναι του ίδιου χρήστη, είναι πολύ περισσότερα, έπρεπε να αντιμετωπίσουμε το πρόβλημα του overfitting.

Το overfitting είναι το πρόβλημα που παρουσιάζεται σε ένα σύστημα κατηγοριοποίησης. Αυτό συμβαίνει όταν δεν κατηγοριοποιεί σωστά λόγω του ότι απομνημονεύει τα δεδομένα τα οποία έχει, με αποτέλεσμα να μην μπορεί να γενικεύει σωστά.

Αυτό το πρόβλημα επιλύθηκε κάνοντας resampling σε όλο το ground truth σε πιο μικρά data sets με διάφορες αναλογίες. Δηλαδή, δημιουργούμε data sets με αναλογίες. Για παράδειγμά, για κάθε μία εγγραφή που αποτελείτε από τα προφίλ του

ίδιου χρήστη βάζουμε είτε 1 είτε 3 είτε 4, 5...10 εγγραφές που περιέχουν προφίλ που δεν ανήκουν στον ίδιο χρήστη.

Για κάθε data set που βγήκε από το resample υπολογίζουμε τις ομοιότητες μέσω προγράμματος υλοποιημένο σε java και Python. Ακολουθεί ψευδοκώδικας της υλοποίησης του κώδικα.

Algorithm 2 Find Distance

```
1: procedure DISTANCE CALCULATION
2:     N ← Data set with users with Facebook and Flickr
3:     for each entry i ∈ N do
4:         Take the name from each profile
5:         Calculate distance using jaro algorithm
6:         Take the location from each profile
7:         Calculate distance using jaro algorithm
8:         Take the profile picture from each profile
9:         Calculate distance using data correlation algorithm
10:        Save results put 1 if the profiles have the same user otherwise put 0
11:    Save results
```

4.3 Ανάλυση Twitter - LinkedIn

Σε αυτή την ενότητα θα παρουσιάσουμε την ανάλυση και τα αποτελέσματα της ανάλυσης πάνω στα δεδομένα τα οποία έχουμε για τους χρήστες που έχουν Twitter και LinkedIn

4.3.1 Χαρακτηριστικά Twitter - LinkedIn

Αρχικά θα βρούμε τα προφίλ για το Twitter και για το LinkedIn για τον ίδιο χρήστη από τα αρχικά δεδομένα τα οποία έχουμε συλλέξει από το FriendFeed.

Αφού συλλέξουμε τα προφίλ που χρειαζόμαστε συνεχίζουμε κάνοντας εξόρυξη χαρακτηριστικών των χρηστών χρησιμοποιώντας τα API των Twitter και LinkedIn. Συνεπώς έχουμε τα χαρακτηριστικά του ίδιου χρήστη και για τα δύο προφίλ που έχει.

Σε αυτά τα δυο κοινωνικά δίκτυα έχουμε περισσότερα χαρακτηριστικά από ότι μας πρόσφερε ο συνδυασμός Facebook-Flickr στην προηγούμενη ανάλυση. Πιο αναλυτικά για το Twitter έχουμε:

1. Όνομα Χρήστη

2. Όνομα Οθόνης

3. Περιγραφή

4. Τοποθεσία

5. Εικόνα προφίλ

Για το LinkedIn:

1. Όνομα Χρήστη

2. Ηλεκτρονική διεύθυνση προφίλ

3. Περίληψη

4. Τίτλος

5. Τοποθεσία

6. Εικόνα προφίλ

Στον πίνακα 4.3 παρουσιάζονται τα χαρακτηριστικά, το είδος και οι πιθανοί αλγόριθμοι σύγκρισης. Αφού τα χαρακτηριστικά μας αποτελούνται κυρίως από συμβολοσειρές θα ακολουθήσουμε τους ίδιους αλγορίθμους που εισηγηθήκαμε για το Facebook και Flickr. Οι συνδυασμοί που θα γίνουν είναι η εξής:

1. Όνομα LinkedIn – Όνομα Twitter

2. Όνομα LinkedIn – Όνομα Οθόνης Twitter

3. Προφίλ url LinkedIn – Όνομα οθόνης Twitter

4. Τοποθεσία LinkedIn – Τοποθεσία Twitter

5. Τίτλος LinkedIn – Περιγραφή Twitter

6. Περίληψη LinkedIn – περιγραφή Twitter

7. Προφίλ εικόνας LinkedIn – Προφίλ εικόνας Twitter

Χαρακτηριστικό	Είδος	Αλγόριθμοι
Όνομα	Συμβολοσειρά	Ngram/Levenshtein/Jaro
Όνομα Οθόνης	Συμβολοσειρά	Ngram/Levenshtein/Jaro
Προφίλ url	Συμβολοσειρά	Ngram/Levenshtein/Jaro
Τοποθεσία	Συμβολοσειρά	Ngram/Levenshtein/Jaro
Τίτλος	Συμβολοσειρά	Ngram/Levenshtein/Jaro
Περίληψη	Συμβολοσειρά	Ngram/Levenshtein/Jaro
Εικόνα προφίλ	Εικόνα	Pixel Wise/Histogram/Data Correlation

Πίνακας 4.3: Δομή, Είδος, Πιθανοί Αλγόριθμοι για Twitter - LinkedIn

Χαρακτηριστικό	Αλγόριθμοι
Όνομα LinkedIn – Όνομα Twitter	Jaro
Όνομα LinkedIn – Όνομα Οθόνης Twitter	Jaro
Προφίλ url LinkedIn – Όνομα οθόνης Twitter	Jaro
Τοποθεσία LinkedIn – Τοποθεσία Twitter	Jaro
Τίτλος LinkedIn – Περιγραφή Twitter	Jaro
Περίληψη LinkedIn – περιγραφή Twitter	Jaro
Προφίλ εικόνας LinkedIn – Προφίλ εικόνας Twitter	Data Correlation

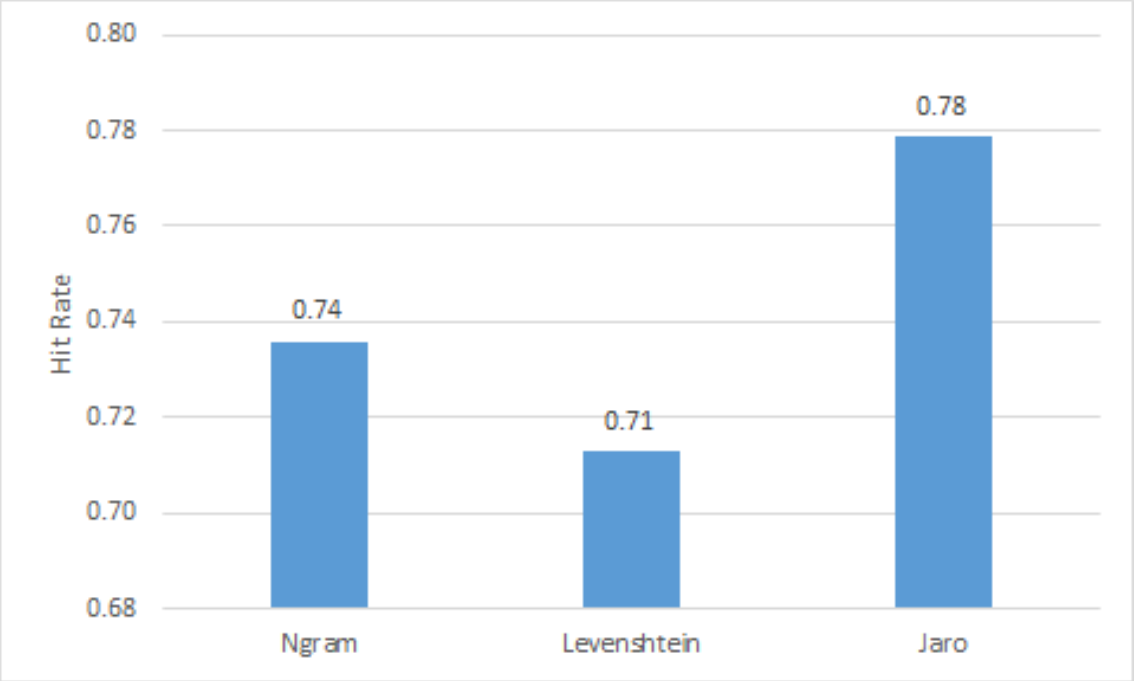
Πίνακας 4.4: Αλγόριθμοι που επιλέχτηκαν για Twitter - LinkedIn

4.3.2 Επιλογή Αλγορίθμων από Ανάλυση

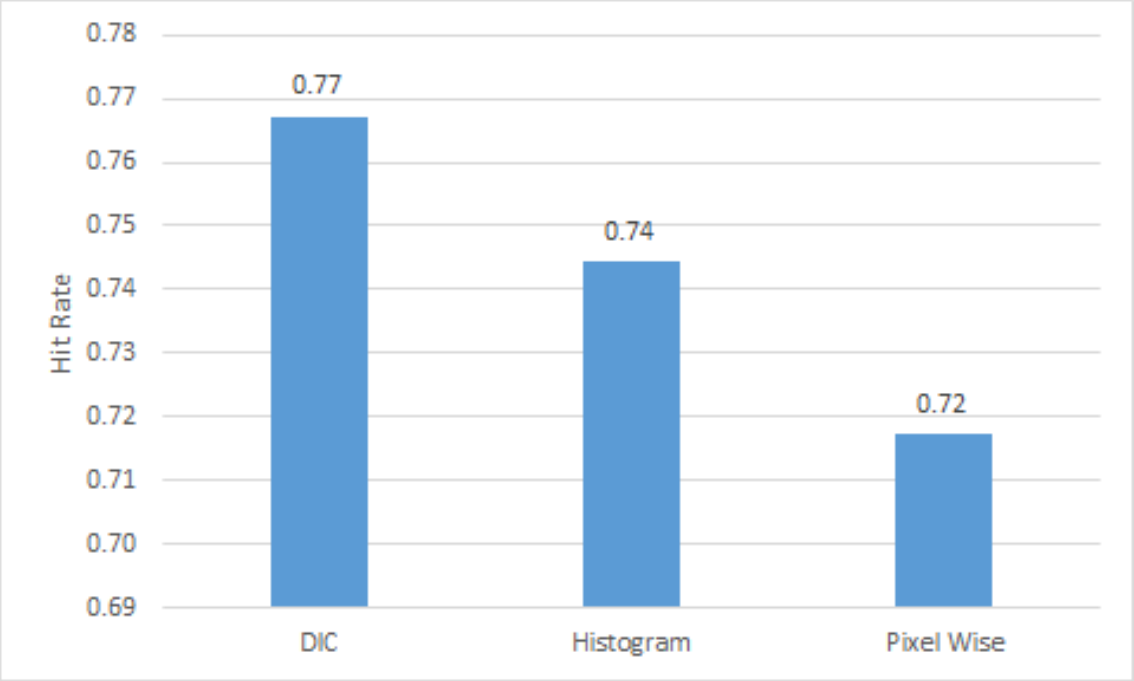
Στη συνέχεια πρέπει να επιλέξουμε τους αλγορίθμους, οι οποίοι θα είναι οι κύριοι αλγόριθμοι για την εύρεση των αποστάσεων/ομοιοτήτων μεταξύ των χαρακτηριστικών που θα χρησιμοποιηθούν για να λύσουν το ερώτημα κατά πόσο δύο προφίλ ανήκουν στον ίδιο χρήστη. Όπως έγινε και στην επιλογή στην προηγούμενη ενότητα.

Ακολουθήσαμε την ίδια τεχνική για την επιλογή αλγορίθμων παίρνοντας τυχαίο δείγμα 700 χρηστών και εφαρμόσαμε πάνω σε αυτό, τους αλγόριθμους για συμβολοσειρες και τους αλγορίθμους για εικόνες. Τα αποτελέσματα φαίνονται στα σχήματα 4.5 και 4.6 για τους string base algorithms και image base algorithms αντίστοιχα

Έχοντας υπόψιν τα αποτελέσματα των γραφημάτων οι προτεινόμενοι αλγόριθμοι παρουσιάζονται στον πίνακα 4.4



Σχήμα 4.5: Hit rate of string base algorithm for Twitter and LinkedIn



Σχήμα 4.6: Hit rate of image base algorithm for Twitter and LinkedIn

4.3.3 Resampling

Μετά την επέκταση των δεδομένων έχουμε συνολικά 37090 εγγραφές από αυτές έχουμε 33606 εγγραφές που είναι no hit και 3483 hit.

Για να καταπολεμήσουμε το overfitting των δεδομένων μας λόγω του ότι έχουμε πολύ περισσότερα no hit σε σχέση με τα hit. Εφαρμόσαμε την τεχνική του resampling όπως την εφαρμόσαμε και στην προηγούμενη ενότητα.

Δηλαδή δημιουργούμε μικρότερα δεδομένα ανάλογα με το λόγο των hit ως προς τα no hit. Έχουμε δημιουργήσει δεδομένα με τις αναλογίες 1-1 και 1-3. Δηλαδή για κάθε μία εγγραφή που είναι σωστή, βάζουμε μία ή τρεις λανθασμένες ανάλογα με την αναλογία. Σύμφωνα με τον πιο κάτω ψευδοκώδικα (Algorithm 3).

Algorithm 3 Find Distance

```
1: procedure DISTANCE CALCULATION
2:     N ← Data set with users with LinkedIn and Twitter
3:     for each entry i ∈ N do
4:         Take the name from each profile
5:         Calculate distance using jaro algorithm
6:         Take the screen name from Twitter and the name from Linked In
7:         Calculate distance using jaro algorithm
8:         Extract user name from Linked In url and screen name from Twitter
9:         Calculate distance using jaro algorithm
10:        Take the summary from Linked In and description from Twitter
11:        Calculate distance using jaro algorithm
12:        Take the headline from Linked In and description from Twitter
13:        Calculate distance using jaro algorithm
14:        Take the location from each profile
15:        Calculate distance using jaro algorithm
16:        Take the profile picture from each profile
17:        Calculate distance using data correlation algorithm
18:        Save results put 1 if the profiles have the same user otherwise put 0
19:    Save results
```

Κεφάλαιο 5

Αλγόριθμος

5.1 Εισαγωγή

Αφού κάναμε την απαιτούμενη ανάλυση στο προηγούμενο κεφάλαιο και καταλήξαμε στα χαρακτηριστικά που θα περιέχονται στον αλγόριθμό μας καθώς και ανάλυση πάνω σε συμπεριφορές των χρηστών, θα συνεχίσουμε με το να βρούμε έναν ευρετικό τρόπο να αξιολογήσουμε κατά πόσο δύο προφίλ ανήκουν στον ίδιο χρήστη.

Για τους λόγους που αναφέρονται στο κεφάλαιο 3, θα προχωρήσουμε στους συνδυασμούς Facebook - Flickr και Twitter - LinkedIn. Δηλαδή αν έχουμε δύο προφίλ Facebook - Flickr ή Twitter - LinkedIn, θα μπορούμε να αποφανθούμε αν τα δύο προφίλ έχουν τον ίδιο χρήστη.

5.2 Logistic Regression

Για το πρόβλημα το οποίο έχουμε, της ταύτισης προφίλ, προτιμήσαμε στην προσέγγιση μας να επιλέξουμε την logistic regression. Ο λόγος της επιλογής αυτού του τρόπου, είναι επειδή ευνοεί τις προβλέψεις συναρτήσεων που αποτελούνται από μεταβλητές. Όπου κάθε μεταβλητή έχει ένα συγκεκριμένο βάρος, διαφορετικό από το κάθε ένα βάρος κάθε μεταβλητής.

Στόχος μας είναι η εύρεση μίας εξίσωσης η οποία να απαντά στο ερώτημα αν δύο προφίλ ανήκουν στον ίδιο χρήστη με μία πιθανότητα που να ανήκει στο κλειστό διάστημα [0,1]. Για αυτό το λόγο σκεφτήκαμε να εφαρμόσουμε Logistic Regression η οποία θα αποτελείται από μία ανεξάρτητη μεταβλητή η οποία είναι δυαδική. Και επίσης θα αποτελείται από μεταβλητές που έχουν κάποιο συντελεστή που θα παρουσιάζει το βάρος και την σημαντικότητα αυτής της μεταβλητής.

$$\log(\frac{p}{1-p}) = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + ... + B_kX_k \quad (5.1)$$

Βασιζόμενος στις διαλέξεις του Δρ. Κωνσταντίνου Φωκιανού [3],παρουσιάζουμε την Logistic Regression όπως φαίνεται στην εξίσωση 5.1. Το p δίνει την πιθανότητα δύο προφίλ να ανήκουν στον ίδιο χρήστη.

Οι συντελεστές παλινδρόμησης εκτιμώνται με τη μέθοδο της μέγιστης πιθανοφάνειας με την υπόθεση ότι η εξαρτημένη μεταβλητή ακολουθεί τη διωνυμική κατανομή. Από την εξίσωση 5.1 το p μπορεί να υπολογιστεί όπως φαίνεται στην εξίσωση 5.2

$$p = \frac{exp(B_0 + B_1X_1 + B_2X_2 + B_3X_3 + ... + B_kX_k)}{1 + exp(B_0 + B_1X_1 + B_2X_2 + B_3X_3 + ... + B_kX_k)} \quad (5.2)$$

5.2.1 Μετρικές Αξιολόγησης

Για να μπορέσουμε να συγκρίνουμε οποιοδήποτε αλγόριθμο ή προσέγγιση χρειαζόμαστε κάποιες μετρικές ως σημεία αναφοράς. Θα χρησιμοποιήσουμε τις μετρικές που φαίνονται στις εξισώσεις 5.3 και 5.4 αντίστοιχα.

$$Precision = \frac{NumberofTruePositives}{(numberofTruePositive + numberofPositiveFalse)} \quad (5.3)$$

$$Recall = \frac{NumberofTruePositives}{(NumberofTruePositives + NumberofFalsenegatives)} \quad (5.4)$$

5.3 Facebook - Flickr

Βασιζόμενοι στην ανάλυση η οποία έγινε για το Facebook - Flickr έχουμε καταλήξει στα χαρακτηριστικά τα οποία θα χρησιμοποιηθούν στον αλγόριθμό μας. Για να έχουμε καλύτερη επίγνωση για το συγκεκριμένο data set στον πίνακα 5.1 παρουσιάζονται τα χαρακτηριστικά που θα χρησιμοποιηθούν στον αλγόριθμο.

5.3.1 Statistical Analysis in Facebook-Flickr

Το data set το οποίο χρησιμοποιήθηκε μετά από πολλές επαναλήψεις περιέχει 12495 εγγραφές. Από αυτές 2496 ανήκουν στην κλάση όπου γνωρίζουμε ότι οι ομοιότητες

Χαρακτηριστικό	Είδος
Όνομα Χρήστη	Συμβολοσειρά
Τοποθεσία Χρήστη	Συμβολοσειρά
Εικόνα Προφίλ Χρήστη	Εικόνα

Πίνακας 5.1: Κύρια χαρακτηριστικά Facebook - Flickr

HIT/NO HIT	Ομοιότητα Ονόματος	Ομοιότητα Τοποθεσίας	Ομοιότητα Εικόνας

Πίνακας 5.2: Μορφή δεδομένων μετά από ανάλυση Facebook - Flickr

που υπολογίστηκαν, ανήκουν σε προφίλ του ίδιου χρήστη ενώ οι υπόλοιπες ανήκουν στην κλάση που δεν ανήκουν στον ίδιο χρήστη. Μετά και το πέρας της ανάλυσης που παρουσιάστηκε στον κεφάλαιο 4 καταφέραμε να χτίσουμε ένα data set όπου έχει την ακόλουθη μορφή όπως στον πίνακα 5.2.

Η στατιστική ανάλυση των δεδομένων έγινε εξολοκλήρου στην R. Συνεπώς έπεται επεξήγηση της χρήσης και των αποτελεσμάτων στην R.

Για την εύρεση της λογιστικής παλινδρόμησης χρησιμοποιήθηκε η πιο κάτω διαδικασία όπως φαίνεται στην εξίσωσή 5.5 Χρησιμοποιώντας αυτή την διαδικασία στην R βγάλαμε κάποια αποτελέσματα τα οποία παρουσιάζονται στον πίνακα 5.3.

$$glm(formula = seen\,name{+}location{+}profilepic, family = binomial(logit), data = ser) \tag{5.5}$$

Βασιζόμενοι στα αποτελέσματα του πίνακα 5.3, παρατηρώντας το p-value για κάθε ένα από τα χαρακτηριστικά μπορούμε να αξιολογήσουμε την σημαντικότητα και την συνεισφορά τους στο μοντέλο το οποίο προτείνουμε για την εύρεση της πιθανότητας δύο προφίλ να ανήκουν στον ίδιο χρήστη.

Το p-value στην στατιστική χρησιμοποιείται για έλεγχο μίας στατιστικής υπόθεσης.

Χαρακτηριστικά	Estimate Std	Error	z value	P value
Intercept	-1.4126	0.1911	-7.392	1.45e-13
Name	-1.8285	0.2022	-9.045	‘2e-16
Location	-0.3342	0.2622	-1.275	0.202
Profile Picture	3.9607	0.3806	10.407	2e-16

Πίνακας 5.3: Αποτελέσματα από ανάλυση Facebook - Flickr

Πριν γίνει κάποιος έλεγχος ορίζουμε ένα κατώφλι, συνήθως 5% ή 1%. Αν το p-value είναι μικρότερο ή ίσο από αυτό το κατώφλι τότε τα δεδομένα μας είναι ασυνεχής και χωρίς συνέπεια. Στην δική μας περίπτωση ορίσαμε το κατώφλι 1% για περισσότερη ακρίβεια και το p-value ήταν μικρότερο.

Την σημαντικότητα και την συνεισφορά κάθε χαρακτηριστικού μπορεί να τα δούμε καλύτερα από την γραφική (βλέπε Σχ. 5.2) που παρουσιάζει την σχετικότητα και την συνεισφορά της κάθε μεταβλητής. Η γραφική είναι αποτέλεσμα ανάλυσης που έγινε στην R.

Συνεπώς είναι προφανές πλέον ότι όνομα και η εικόνα προφίλ του χρήστη παίζουν σημαντικό ρόλο στην πρόβλεψη κατά πόσο δύο προφίλ να ανήκουν στον ίδιο χρήστη.

Χωρίς να εφαρμόσουμε οποιαδήποτε προσέγγιση για training set και test set επιχειρήσαμε να δούμε πια θα ήταν τα αποτελέσματα πάνω σε όλο το σύνολο των δεδομένων μας.

Έτσι βρήκαμε τα predicted values[1] όπως φαίνεται στο σχήμα 5.1, όπου παρουσιάζονται τα αποτελέσματα για κάθε μία από τις εγγραφές που έχουμε στο data set μας.

Πιο αναλυτικά, το σχήμα περιέχει όλες τις εγγραφές που έχουμε στο δείγμα μας. Για κάθε μία από τις εγγραφές εκτελείτε η logistic regression και το αποτέλεσμα τις κάθε εγγραφής, παρουσιάζεται στην γραφική.

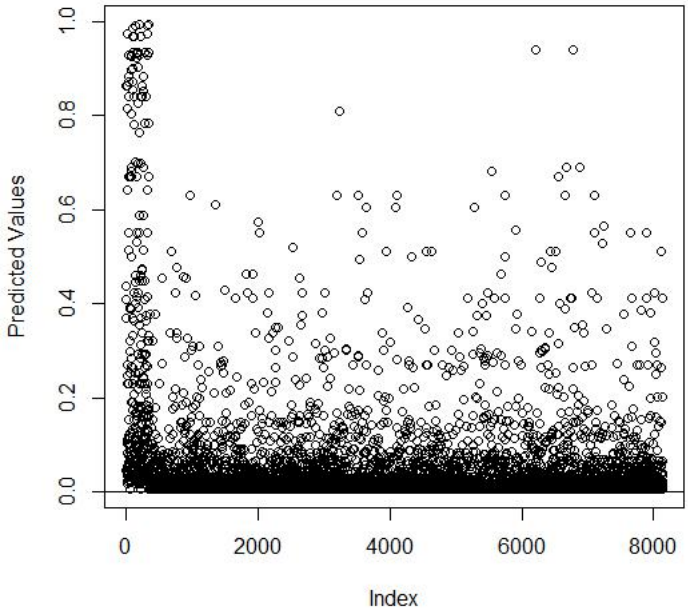
Παρατηρούμε ότι τα περισσότερα αποτελέσματα κυμαίνονται από το (0,0.2]. Βλέποντας αυτήν την συμπεριφορά στα αποτελέσματα μας αντιλαμβανόμαστε ότι το μοντέλο μας δεν θα έχει καλά αποτελέσματα. Αυτό φαίνεται και από των αριθμό των false negatives, οι πρώτες 2000 εγγραφές που γνωρίζουμε ότι τα προφίλ τους ανήκουν στον ίδιο χρήστη έχουν χαμηλή πιθανότητα ενώ στην πραγματικότητα θα έπρεπε να έχουν ψηλή πιθανότητα.

Μια γενική παρατήρηση που γίνεται αντιληπτή αμέσως, είναι ο μικρός αριθμός από χαρακτηριστικά τα οποία έχουμε στην κατοχή μας. Αυτό μας επιβάλλεται λόγω των διάφορων πολιτικών που ακολουθούν τα κοινωνικά δίκτυα και ειδικά από το Facebook.

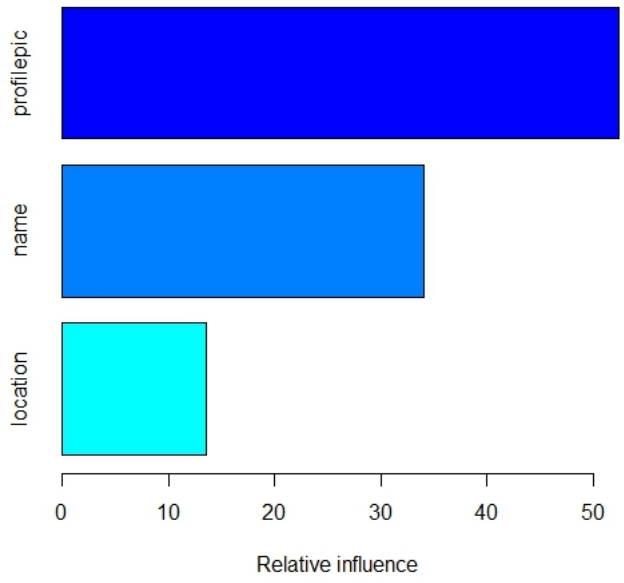
Το Facebook είναι πολύ ευαίσθητο σε θέματα που έχουν να κάνουν με την διαχείριση και παρουσίαση των προσωπικών δεδομένων των χρηστών του με αποτέλεσμα να μας επιτρέπει να δούμε πολύ λίγη πληροφορία.

Έχοντας υπόψιν το πιο πάνω πρόβλημα πρέπει να σκεφτούμε ένα τρόπο όπου τα δεδομένα μας πρέπει να γίνουν train. Δηλαδή ένα τρόπο όπου να μπορεί να βελτιώσει τα προηγούμενα αποτελέσματα (σχήμα 5.1).

[1] Αποτέλεσμα από logistic regression



Σχήμα 5.1: Πιθανότητα να ανήκουν τα δύο προφίλ στον ίδιο χρήστη.



Σχήμα 5.2: Σχετικότητα και συνεισφορά της κάθε μεταβλητής

Χαρακτηριστικά	Βάρη
Intercept	-1.4118
Name	-1.88741
Location	-0.4171
Profile Picture	4,0508

Πίνακας 5.4: Εκπαιδευμένα Βάρη

5.3.2 Δεδομένα Εκπαίδευσης και Ελέγχου

Αφού αναγνωρίσαμε το πρόβλημα με τα λιγοστά χαρακτηριστικά που έχουμε στην διάθεσή μας, χωρίσαμε τα δεδομένα μας σε δεδομένα ελέγχου και δεδομένα εκπαίδευσης με τον κανόνα 70 – 30. Αυτός ο τρόπος είναι ιδιαίτερα διαδεδομένος στον χώρο της τεχνητής νοημοσύνης και του machine learning. Θα τον χρησιμοποιήσουμε για να δούμε κατά πόσον ο αλγόριθμος μας εκπαιδεύεται σωστά και αν γενικεύει σωστά.

Πρώτα, διαιρούμε τα δεδομένα που ανήκουν σε ολόκληρο το δείγμα σε 70% και 30%, στη συνέχεια παίρνουμε το 70% και τρέχουμε πάνω σε αυτό έναν συγκεκριμένο αλγόριθμο, στην δική μας περίπτωση τρέχουμε την logistic regression, μετά το πέρας αυτού του αλγορίθμου συλλέγουμε τα βάρη τα οποία έχουν βγει από την ανάλυση της logistic regression. Τα βάρη παρουσιάζουν την σημαντικότητα και την συνεισφορά κάθε μεταβλητής.

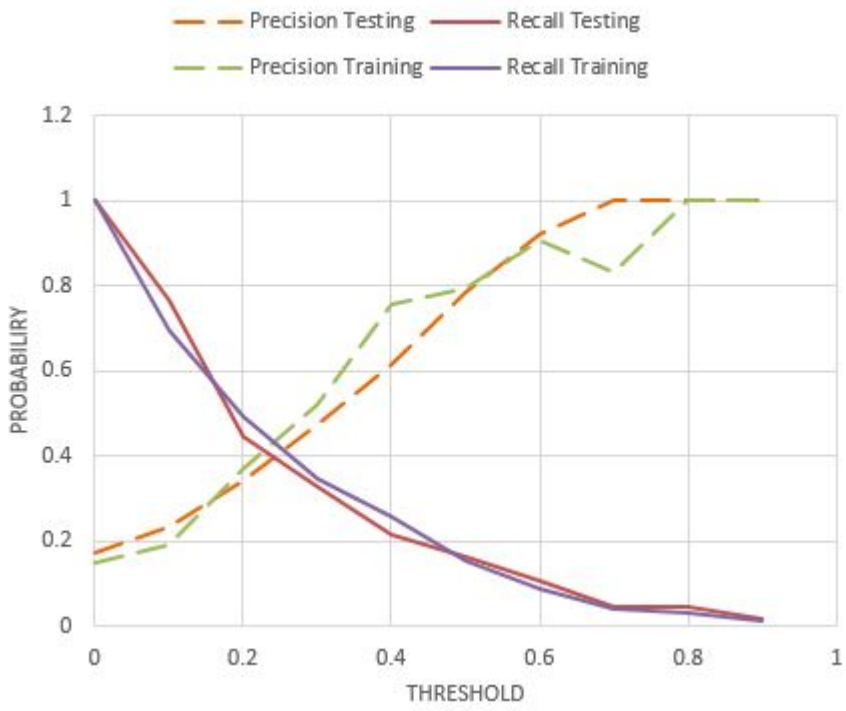
Στην συνέχεια, αφού πάρουμε αυτά τα βάρη (βλέπε πίνακα 5.4) , στο δείγμα που έχει το 30% του συνολικού δείγματος τρέχουμε την εξίσωση η οποία έχει παραχθεί από την logistic regression χρησιμοποιώντας αυτά τα βάρη.

Η πιθανότητα που βγαίνει από την εξίσωση μας παίρνει τιμές του διαστήματος [0,1]. Συνεπώς χρειαζόμαστε ένα σημείο αναφοράς που θα καθορίζει κατά πόσο δύο προφίλ έχουν τον ίδιο χρήστη.

Για αυτό το λόγο εισάγαμε ένα κατώφλι το οποίο θα καθορίζει αν δύο προφίλ έχουν τον ίδιο χρήστη. Αν η πιθανότητα που θα βγει από την εξίσωση είναι μικρότερη από το κατώφλι τότε τα προφίλ δεν ανήκουν στον ίδιο χρήστη αν είναι μεγαλύτερη ή ίση τότε τα δύο προφίλ ανήκουν στον ίδιο χρήστη.

Εισάγαμε ένα κατώφλι το οποίο παίρνει τιμές από 0,1 μέχρι 0,9 και τρέξαμε την εξίσωση μας με όλα τα πιθανά κατώφλια, για να μπορέσουμε να βρούμε το βέλτιστο κατώφλι. Παρατηρούμε ότι όσο το κατώφλι αυξάνεται το precision αυξάνεται ενώ το recall μειώνεται (βλέπε Σχ. 5.3).

Λόγω του ότι το κατώφλι αυξάνεται, η πιθανότητα που βγαίνει από τον αλγόριθμο δεν είναι τόσο μεγάλη που να ξεπερνά το κατώφλι, με αποτέλεσμα να απαντάει συνέχεια



Σχήμα 5.3: Precision και Recall για Δεδομένα Ελέγχου/Εκπαίδευσης

Χαρακτηριστικά	Βάρη
Intercept	-2.723
Name	-2.189
Profile Picture	7.674

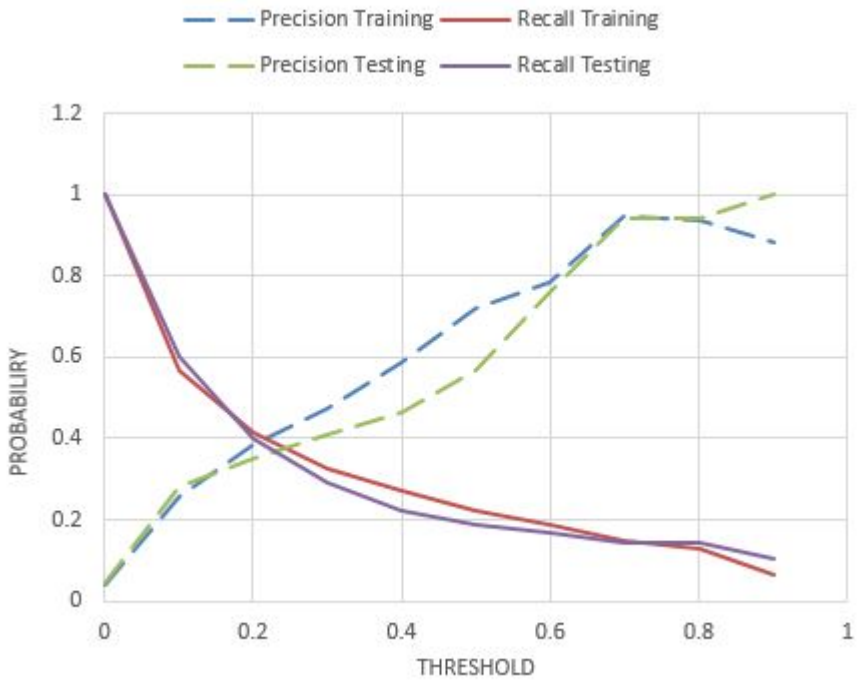
Πίνακας 5.5: Εκπαιδευμένα Βάρη χωρίς την τοποθεσία

'ΟΧΙ' στο ερώτημα κατά πόσο τα δύο προφίλ ανήκουν στον ίδιο χρήστη.

Συνεπώς το recall, που είναι η μετρική που αξιολογεί όλες τις απαντήσεις του αλγορίθμου, μειώνεται αισθητά σε αντίθεση με το precision, που είναι η μετρική που αξιολογεί τις ορθές απαντήσεις, να αυξάνεται.

Αφού είδαμε ότι δεν έχουμε τόσο καλά αποτελέσματα από αυτή την ανάλυση, βασιζόμενοι στο σχήμα 5.2, που παρουσιάζει την σημαντικότητα των χαρακτηριστικών, αποφασίσαμε να αφαιρέσουμε την τοποθεσία από την ανάλυση, για να δούμε αν θα πάρουμε καλύτερα αποτελέσματα από πριν. Μετά την αφαίρεση της τοποθεσίας πρέπει να γίνει από την αρχή η ανάλυση της logistic regression για να βρούμε την νέα εξίσωση και τα νέα βάρη που θα παραχθούν(βλέπε πίνακα 5.5).

Με την αφαίρεση της τοποθεσίας, τα δεδομένα μας έχουν αυξηθεί σε 5667 για δεδομένα εκπαίδευσης και 2468 σε δεδομένα ελέγχου. Όπως βλέπουμε και στη γραφική σχήμα 5.4 δεν παρατηρείτε κάποια βελτίωση όσο αφορά το precision και recall σε



Σχήμα 5.4: Precision και Recall για Δεδομένα Ελέγχου/Εκπαίδευσης χωρίς την τοποθεσία του χρήστη

σχέση με πριν.

Η τελική εξίσωση που δίνει την πιθανότητα κατά πόσον δύο προφίλ να ανήκουν στον ίδιο χρήστη παρουσιάζεται στην εξίσωση 5.4.

$$p = (-1.4118) + (-1.88741) * X_0 + (-0.4171) * X_1 + (4.0508) * X_2 \tag{5.6}$$

Όπου:

X0

Είναι η ομοιότητα των ονομάτων των χρηστών.

X1

Είναι η ομοιότητα των τοποθεσιών των χρηστών.

X2

Είναι η ομοιότητα των εικόνων των χρηστών.

5.4 LinkedIn - Twitter

Βασιζόμενοι στην ανάλυση η οποία έγινε για το LinkedIn - Twitter έχουμε καταλήξει στα χαρακτηριστικά τα οποία θα χρησιμοποιηθούν στον αλγόριθμό μας.

Χαρακτηριστικό
Όνομα LinkedIn – Όνομα Twitter
Όνομα LinkedIn – Όνομα Οθόνης Twitter
Προφίλ url LinkedIn – Όνομα οθόνης Twitter
Τοποθεσία LinkedIn – Τοποθεσία Twitter
Τίτλος LinkedIn – Περιγραφή Twitter
Περίληψη LinkedIn – περιγραφή Twitter
Προφίλ εικόνας LinkedIn – Προφίλ εικόνας Twitter

Πίνακας 5.6: Χαρακτηριστικά για Twitter - LinkedIn

5.4.1 Statistical Analysis in LinkedIn - Twitter

Αφού βρήκαμε τα αποτελέσματα των μεθόδων προχωρήσαμε στην ανάλυση για να εξάγουμε χρήσιμες πληροφορίες για να κατασκευάσουμε τον αλγόριθμό μας για το Twitter και το LinkedIn. Για ευκολότερη κατανόηση έχουμε μετονομάσει τις μεταβλητές:

X0

Όνομα LinkedIn – Όνομα Twitter

X1

Όνομα LinkedIn – Όνομα Οθόνης Twitter

X2

Προφίλ url LinkedIn – Όνομα οθόνης Twitter

X3

Περίληψη LinkedIn – περιγραφή Twitter

X4

Τίτλος LinkedIn – Περιγραφή Twitter

X5

Τοποθεσία LinkedIn – Τοποθεσία Twitter

Χαρακτηριστικά	Estimate Std	Error	z value	P value
Intercept	-7,929135	0,30482	-26,013	2e-16
X0	-1.634419	0.215371	7.589	3.23e-14
X1	0.982264	0.198747	4.942	7.72e-07
X2	2.679944	0.484000	5.537	3.08e-08
X3	-0.008839	0.224166	-0.039	0.969
X4	3.007144	0.233870	12.858	2e-16
X5	3.774730	0.160651	23.496	2e-16
X6	4.485379	0.218932	20.488	2e-16

Πίνακας 5.7: Αποτελέσματα από ανάλυση Twitter - LinkedIn

X6

Προφίλ εικόνας LinkedIn – Προφίλ εικόνας Twitter

Για την εύρεση της logistic regression χρησιμοποιήθηκε η πιο κάτω διαδικασία (5.7) με τα αντίστοιχα αποτελέσματα όπως φαίνονται στον πίνακα 5.7.

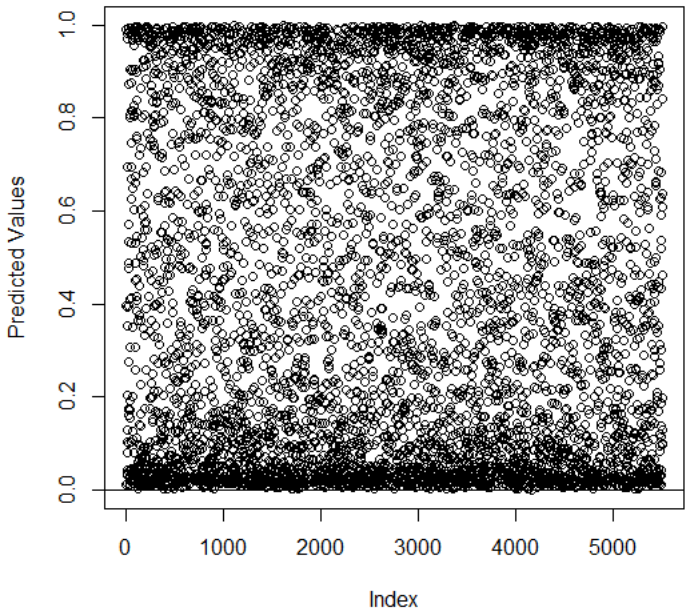
$$glm(formula = class\ X0 + X1 + X2 + X3 + X4 + X5 + X6 , family = binomial(logit), data = ser) \quad (5.7)$$

Βασιζόμενοι στα αποτελέσματα του πίνακα 5.7 και παρατηρώντας το p-value για κάθε ένα από τα χαρακτηριστικά μπορούμε να αξιολογήσουμε την σημαντικότητα και την συνεισφορά τους στο μοντέλο το οποίο προτείνουμε για την εύρεση της πιθανότητας δύο προφίλ να ανήκουν στον ίδιο χρήστη.

Αυτό μπορεί να το δούμε καλύτερα από την γραφική (Σχ.5.6) που παρουσιάζει την σχετικότητα και την συνεισφορά της κάθε μεταβλητής. Η γραφική είναι αποτέλεσμα ανάλυσης που έγινε στην R. Συνεπώς είναι προφανές πλέον ότι οι μεταβλητές X5, X6, X4, X0, X1 είναι οι πιο σημαντικές και παίζουν σημαντικό ρόλο στην πρόβλεψη κατά πόσο δύο προφίλ να ανήκουν στον ίδιο χρήστη.

Χωρίς να εφαρμόσουμε οποιαδήποτε προσέγγιση για training and test set επιχειρήσαμε να δούμε πια θα ήταν τα αποτελέσματα πάνω σε όλο το σύνολο των δεδομένων μας.

Στη συνέχεια παρουσιάζουμε τα predicted values σε γραφική τρέχοντας τον αλγόριθμο μας σε όλα τα δεδομένα (Σχήμα 5.5). Πιο συγκεκριμένα για κάθε μία από τις εγγραφές εκτελείται η φόρμουλα της logistic regression και βγαίνει για κάθε εγγραφή η πιθανότητα δύο προφίλ να ανήκουν στον ίδιο χρήστη.



Σχήμα 5.5: Πιθανότητα να ανήκουν τα δύο προφίλ στον ίδιο χρήστη Twitter - LinkedIn.

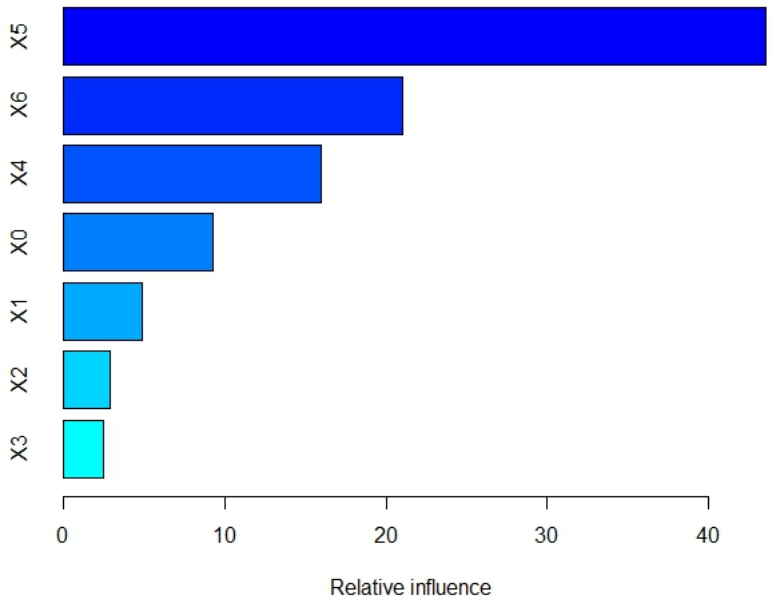
Όπως φαίνεται και από το σχήμα παρατηρούμε ότι οι πιθανότητες που παρουσιάζονται κατανέμονται σε όλο το φάσμα των δεδομένων μας. Σαν μια πρώτη ένδειξη, μπορούμε να πούμε ότι θα έχουμε καλύτερα αποτελέσματα σε σχέση με το Facebook - Flickr.

5.4.2 Δεδομένα Εκπαίδευσης και Ελέγχου

Για την ανάλυση των δεδομένων μας έχουμε χρησιμοποιήσει την μέθοδο 70-30 με 70% των δεδομένων μας να είναι δεδομένα εκπαίδευσης και 30% να είναι δεδομένα ελέγχου.

Πρώτα, διαιρούμε τα δεδομένα που ανήκουν σε ολόκληρο το δείγμα σε 70% και 30%, στη συνέχεια παίρνουμε το 70% και τρέχουμε πάνω σε αυτό έναν συγκεκριμένο αλγόριθμο, στην δική μας περίπτωση τρέχουμε την logistic regression, μετά το πέρας αυτού του αλγορίθμου συλλέγουμε βάρη τα οποία έχουν βγει από την ανάλυση της logistic regression. Τα βάρη παρουσιάζουν την σημαντικότητα και την συνεισφορά κάθε μεταβλητής.

Στην συνέχεια, αφού πάρουμε αυτά τα βέλτιστα βάρη (βλέπε πίνακα 5.8), στο δείγμα που έχει το 30% του συνολικού δείγματος τρέχουμε την εξίσωση η οποία έχει παραχθεί από την logistic regression χρησιμοποιώντας αυτά τα βάρη. Τα αποτελέσματα αυτής της διαδικασίας, παρουσιάζονται στον πίνακα 5.9.



Σχήμα 5.6: Σχετικότητα και συνεισφορά της κάθε μεταβλητής Twitter - LinkedIn

Χαρακτηριστικά	Βάρη
Intercept	-8,01125
X0	1,8247
X1	0,87442
X2	2,56482
X3	0,09911
X4	3,02653
X5	3,75589
X6	4,44946

Πίνακας 5.8: Βέλτιστα Βάρη Twitter - LinkedIn

Κατώφλι	Accuracy	Precision	Recall
0.0	0.42	0.42	1.0
0.1	0.72	0.61	0.96
0.2	0.78	0.68	0.92
0.3	0.81	0.73	0.88
0.4	0.84	0.79	0.85
0.5	0.85	0.85	0.79
0.6	0.86	0.90	0.76
0.7	0.84	0.93	0.69
0.8	0.81	0.94	0.59
0.9	0.75	0.96	0.44

Πίνακας 5.9: Αποτελέσματα Twitter - LinkedIn

Έχοντας υπόψιν το σχήμα 5.6, που παρουσιάζει την συνεισφορά του κάθε χαρακτηριστικού, επιχειρήσαμε να αφαιρέσουμε το χαρακτηριστικό X3 διότι από την ανάλυση που έγινε δεν συνεισφέρει πάρα πολύ στην απόφαση κατά πόσο δύο προφίλ ανήκουν στον ίδιο χρήστη, για να ελέγξουμε αν προκύψουν καλύτερα αποτελέσματα. Τα αποτελέσματα παρουσιάζονται στον πίνακα 5.10. Παρατηρούμε ότι παίρνουμε σχεδόν παρόμοια αποτελέσματα σε σχέση με πριν.

Η τελική εξίσωση που εισηγούμαστε για το Twitter - LinkedIn δίνεται στη εξίσωση 5.8

$$p = (-8.01125) + (-1.8247) * X_0 + (-0.87442) * X_1 + (2.56482) * X_2 + (0.09911) * X_3 + (3.02653) * X_4 + (3.75589) * X_5 + (4.44946) * X_6 \quad (5.8)$$

X0

Όνομα LinkedIn – Όνομα Twitter

X1

Όνομα LinkedIn – Όνομα Οθόνης Twitter

X2

Προφίλ url LinkedIn – Όνομα οθόνης Twitter

X3

Περίληψη LinkedIn – περιγραφή Twitter

Κατώφλι	Accuracy	Precision	Recall
0.0	0.42	0.42	1.0
0.1	0.71	0.60	0.96
0.2	0.77	0.67	0.92
0.3	0.81	0.73	0.89
0.4	0.83	0.78	0.85
0.5	0.85	0.84	0.80
0.6	0.86	0.89	0.77
0.7	0.84	0.93	0.70
0.8	0.82	0.94	0.61
0.9	0.75	0.96	0.45

Πίνακας 5.10: Αποτελέσματα χωρίς το X3 Twitter - LinkedIn

X4

Τίτλος LinkedIn – Περιγραφή Twitter

X5

Τοποθεσία LinkedIn – Τοποθεσία Twitter

X6

Προφίλ εικόνας LinkedIn – Προφίλ εικόνας Twitter

Κεφάλαιο 6

Evaluation

6.1 Εισαγωγή

Μετά από την εισήγηση μας με την logistic regression και τα αποτελέσματα που παρουσιάζονται στο κεφάλαιο 5. Θα πρέπει να παρουσιάσουμε αν η δική μας προσέγγιση βγάζει καλύτερα ή χειρότερα αποτελέσματα σε σχέση με υφιστάμενους αλγορίθμους.

6.2 Αλγόριθμοι Σύγκρισης

6.2.1 Naïve Bayes

Οι ταξηνομητές Naïve Bayes [15] βασίζονται στο να εφαρμόζουν το θεώρημα του Bayes έχοντας μεγάλη ανεξαρτησία μεταξύ των χαρακτηριστικών. Επίσης, είναι εξαιρετικά επεκτάσιμη, το μόνο που χρειάζεται είναι μία σειρά από γραμμικούς παραμέτρους σε ένα πρόβλημα μάθησης.

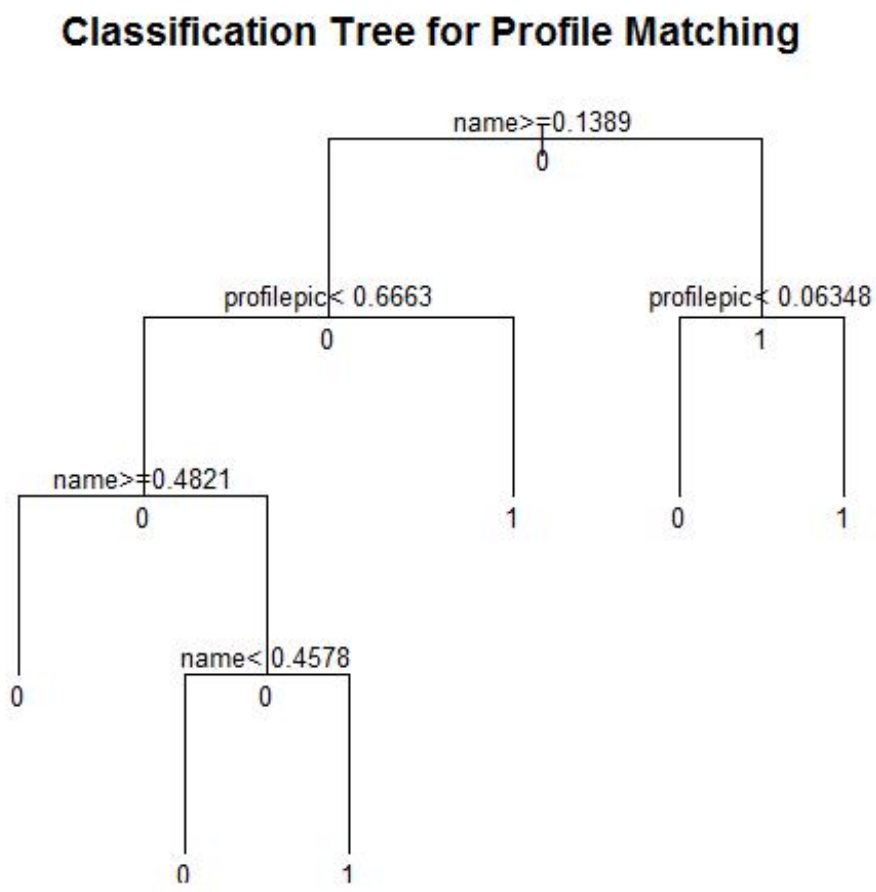
6.2.2 Decision Tree

Το Decision Tree [11] είναι ένα εργαλείο υποστήριξης αποφάσεων που χρησιμοποιεί ένα δέντρο που μοιάζει με γραφική παράσταση ή μοντέλο αποφάσεων με πιθανές συνέπειές.

6.3 Facebook - Flickr

6.3.1 Προτεινόμενος Αλγόριθμος

Τρέχοντας την δική μας προσέγγιση,για τον συνδυασμό του Facebook και του Flickr, βγάζουμε ποσοστό επιτυχίας γύρο στο 0,58.



Σχήμα 6.1: Δέντρο Απόφασης Facebook - Flickr

Φυσικά το αποτέλεσμα πάντα εξαρτώνται από τα δεδομένα τα οποία θα επεξεργαστούμε. Τα αποτελέσματα που βρήκαμε αναμένονταν να είναι χαμηλά λόγω του ότι δεν έχουμε αρκετές μεταβλητές/χαρακτηριστικά για να μπορέσουμε να βγάλουμε πιο ακριβής αποτελέσματα.

6.3.2 Αποτελέσματα από Naïve Bayes και Decision Tree

Με βάση το δένδρο απόφασης (βλέπε Σχ. 6.1), όπου η υλοποίηση του έγινε με τη βοήθεια της προγραμματιστικής γλώσσας Java, είχε όσο αφορά το Precision 0,400 και 0,100 για το Recall.

Για τα αποτελέσματα του Bayes η εύρεση των αποτελεσμάτων έγινε στην R με 0,400 για Precision και 0,500 για το Recall.

Αλγόριθμοι	Precision	Recall
Naïve Bayes	0.400	0.500
Decision Tree	0.400	0.100
Our approach	0.58	0.55

Πίνακας 6.1: Final Results for Flickr and Facebook

Φαίνεται πως καταφέραμε να δημιουργήσουμε μία προσέγγιση που επιστρέφει πιο ακριβής αποτελέσματα σε σχέση με τις υπόλοιπες μεθόδους (πίνακας 6.1).

Από την άλλη, θα μπορούσαμε να πάρουμε ακόμη καλύτερα αποτελέσματα αλλά για να γίνει αυτό θα πρέπει να συλλέξουμε και άλλα χαρακτηριστικά από τους χρήστες του Facebook.

Αυτό είναι αδύνατον με συμβατικούς τρόπους (π.χ. Api), συνεπώς η καλύτερη πρόβλεψη που μπορούμε να κάνουμε με τα υπάρχον δεδομένα μας είναι γύρο στο 58%.

6.4 Twitter - LinkedIn

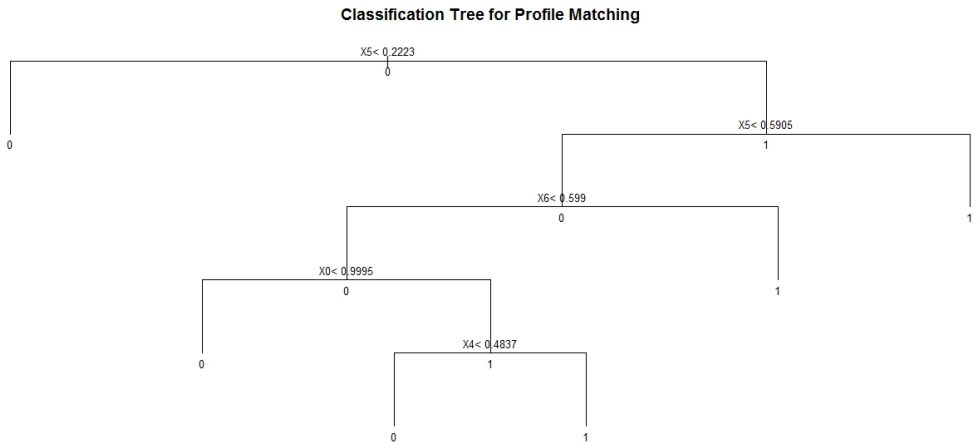
6.4.1 Προτεινόμενος Αλγόριθμος

Τρέχοντας τον αλγόριθμο που εισηγούμαστε, πάνω στα δεδομένα ελέγχου παρατηρούμε πως παίρνουμε πολύ καλά αποτελέσματα σε σχέση με το Facebook και Flickr αυτό οφείλεται στο ότι έχουμε πολύ περισσότερα χαρακτηριστικά για να μπορέσουμε να κάνουμε εφαρμογή του αλγορίθμου μας.

Για κατώφλι 0,6 παίρνουμε Precision 0,90 και Recall 0,76 και για κατώφλι 0,5 παίρνουμε Precision 0,85 και Recall 0, 79.

Στα αποτελέσματα όπου αφαιρέσαμε το χαρακτηριστικό X3[1] διότι από την ανάλυση που έγινε δεν συνεισφέρει πάρα πολύ στην απόφαση κατά πόσο δύο προφίλ ανήκουν στον ίδιο χρήστη για να ελέγξουμε αν προκύψουν καλύτερα αποτελέσματα. Παρατηρούμε ότι παίρνουμε κάπως παρόμοια αποτελέσματα σε σχέση με πριν.

Από τον συνδυασμό του Twitter και LinkedIn έχουμε ένα ποσοστό επιτυχίας γύρο στο 76% με 90%. Φυσικά τα αποτελέσματα εξαρτώνται πάντα από τα δεδομένα τα οποία θα επεξεργαστούμε.



Σχήμα 6.2: Δέντρο Απόφασης Twitter - LinkedIn

Μεθόδοι	Precision	Recall
Naïve Bayes	0,70	0,86
Decision Tree	0,55	0,65
Our Approach	0,84	0,80

Πίνακας 6.2: Final Results for Twitter and LinkedIn

6.4.2 Αποτελέσματα από Naïve Bayes και Decision Tree

Με βάση το δένδρο απόφασης (βλέπε Σχ. 6.2), όπου η υλοποίηση του έγινε με τη βοήθεια της προγραμματιστικής γλώσσας Java, είχε όσο αφορά το Precision είναι 0,55 και 0,65 για το Recall.

Για τα αποτελέσματα του Bayes η εύρεση των αποτελεσμάτων έγινε στην R με 0,70 για Precision και 0,86 για το Recall(πίνακας 6.2).

6.5 Συμπεράσματα

Όπως φαίνεται και από τις συγκρίσεις που έχουμε παρουσιάσει φαίνεται πως έχουμε δημιουργήσει μία προσέγγιση αρκετά ακριβείς και πολύ υποσχόμενη, όσο αφορά τα αποτελέσματα των τετριμμένων τεχνικών.

Πιο αναλυτικά έχουμε δείξει στις προηγούμενες ενότητες τα αποτελέσματα διάφορων τεχνικών σε σχέση με την δική μας προσέγγιση. Στον πίνακα 6.3 παρουσιάζονται

[1]Περίληψη LinkedIn – περιγραφή Twitter

Συνδυασμοί	Precision	Recall
Twitter - LinkedIn	89%	79%
Facebook - Flickr	58%	55%

Πίνακας 6.3: Final Results for Twitter and LinkedIn

τα αποτελέσματα που έχουμε βγάλει για τους δύο συνδυασμούς που επιλέξαμε.

Κεφάλαιο 7

Web App

7.1 Εισαγωγή

Στα προηγούμενα κεφάλαια παρουσιάστηκε η δική μας προσέγγιση που έρχεται να προσφέρει λύση στο πρόβλημα που προσπαθούμε να λύσουμε. Το πρόβλημα της ταύτισης των προφίλ ενός χρήστη είναι η αρχή για επίλυση ακόμη πιο μεγάλων προβλημάτων όπως αυτών που παρουσιάστηκαν στο κεφάλαιο 1.

Αφού παρουσιάσαμε τα αποτελέσματα και δείξαμε ότι η προσέγγιση μας είναι καλύτερη σε σχέση με τετριμμένες μεθόδους, θα πρέπει να χτίσουμε ένα εργαλείο που θα βασίζεται πάνω στην προσέγγιση μας και θα προσφέρει αυτή την λειτουργία στους χρήστες.

7.2 Wally

Σε αυτή την ενότητα θα παρουσιάσουμε το εργαλείο το οποίο έχουμε δημιουργήσει, τον Wally[1]. Το εργαλείο αυτό δίνει την δυνατότητα σε έναν χρήστη να βρίσκει τα προφίλ ενός άλλου χρήστη εισάγοντας μόνο ένα profile url ενός κοινωνικού δικτύου οποιουδήποτε άλλου χρήστη.

7.2.1 Πρόβλημα

Πέραν του προβλήματος της ταύτισης των διάφορων προφίλ κάποιου χρήστη στα κοινωνικά δίκτυα ο Wally έρχεται να λύσει και το πρόβλημα του χρόνου. Μειώνει στο μέγιστο τον χρόνο που σπαταλάτε πάνω σε μία αναζήτηση πάνω στα κοινωνικά δίκτυα.

Το εργαλείο εφαρμόζει τον αλγόριθμο που έχουμε αναφέρει στα προηγούμενα κεφάλαια και παρουσιάζει τα προφίλ τα οποία είναι πιθανόν να ανήκουν στον ίδιο χρήστη

[1] http://www.wideo.co/view/9332911430056902552-wally?from=cp

Σχήμα 7.1: Wally©2015

με μία πιθανότητα.

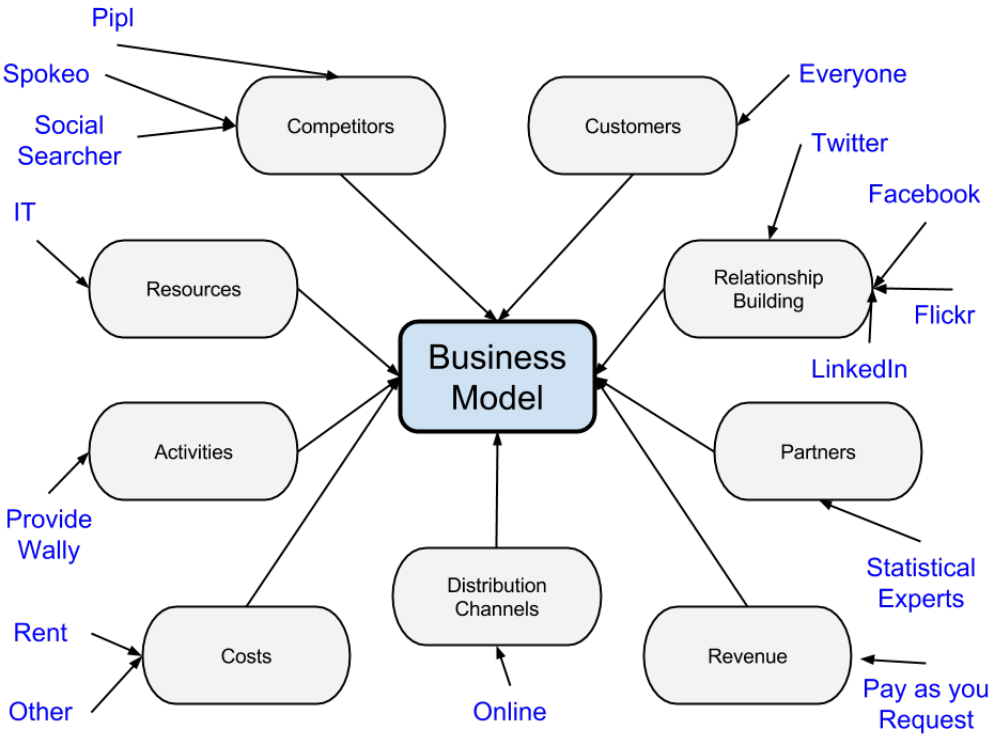
7.2.2 Business Model

Θα παρουσιάσουμε το επιχειρηματικό μοντέλο(βλέπε Σχ. 7.2) το οποίο εισηγούμαστε για τον Wally. Αρχικά θα παρουσιάσουμε τους ανταγωνιστές οι οποίοι παρέχουν παρόμοια λειτουργία με αυτήν του Wally. Οι εφαρμογές Pipl, Spokeo, Social Searcher παρέχουν την ίδια λειτουργία με την διαφορά ότι εμείς ζητάμε από τον χρήστη να εισάγει ένα profile url ενός κοινωνικού δικτύου, ενώ οι ανταγωνιστές μας, δίνουν την δυνατότητα να ψάξει ένας χρήστης με βάση το όνομα ή και άλλα χαρακτηριστικά. Εμείς έχοντας το profile url παίρνουμε όλα τα χαρακτηριστικά υπόψιν μας από την αρχή.

Τα αρχικά κόστη της εφαρμογής θα είναι μηδενικά μιας και δεν χρειάζεται κάποια ιδιαίτερη υποδομή για να τρέξει. Το κύριο έσοδο θα είναι τα αποτελέσματα έναντι αμοιβής.Αρχικά η εφαρμογή θα είναι δωρεάν και μετά απο κάποιο συγκεκριμένο χρονικό διάστημα θα υπάρχει μία χρέωση για κάθε request που κάνει ένας χρήστης. Όσο αφορά τα υπόλοιπα μέρη του business model παρουσιάζονται στο σχήμα 7.2.

7.2.3 Underlying Magic

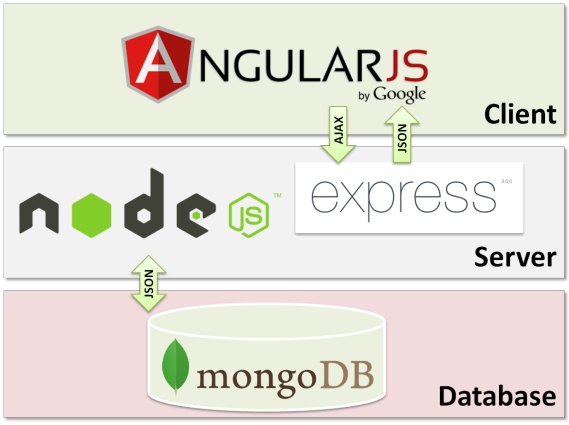
Μετά την παρουσίασή του προβλήματος που προσπαθεί να λύση ο Wally καθώς και την ανάλυση του επιχειρηματικού πλάνου, προχωρήσαμε στην υλοποίηση της εφαρμογής.



Σχήμα 7.2: Business Model Wally

Για να καταλήξουμε σε μία συγκεκριμένη τεχνολογία προχωρήσαμε σε μία ενδελεχή έρευνα πάνω σε διάφορες τεχνολογίες.

Για την υλοποίηση της εφαρμογής χρησιμοποιήθηκε η αρχιτεκτονική MEAN Stack[2] η οποία παρέχει μία πληθώρα από τεχνολογίες και ευκολίες τόσο στην υλοποίηση όσο και στην απόδοση. Πιο αναλυτικά θα χρησιμοποιήσουμε για front end την Anqular JS[3] για back end την Node JS[4] την Express JS[5] η οποία είναι ένα web framewotk for Node JS και για βάση δεδομένων θα χρησιμοποιήσουμε την την Mongo Db (NonSql Database)[6]



Σχήμα 7.3: Mean Stack Architecture

[2]http://mean.io
[3]https://angularjs.org/
[4]https://nodejs.org/
[5]http://expressjs.com/
[6]https://www.mongodb.org/

Οι λόγοι που μας ώθησαν στο να επιλέξουμε το MEAN Stack είναι οι ακόλουθοι:

1. Σου δίνει την δυνατότητα να γράψεις ολόκληρο τον κώδικα σου σε JavaScript, απο client to server
2. Χρησιμοποιεί ένα από τα πιο αποδοτικά design patterns το Model-View-Controller, δίνοντας σου ευκολία στη διαχείριση του κώδικά σου.
3. Όλα τα συστατικά του MEAN είναι open-source projects
4. Συνεχώς υπάρχουν αναβάθμισης και νέες εκδόσεις

AngularJS

Η AngularJS είναι ένα front-end framework το οποίο υλοποιήθηκε από την Google, και σε βοηθά μα χτίσεις την εφαρμογή σου σε modules. Επίσης σε βοηθά να χτίσεις web apps με τον ελάχιστο αριθμό από γραμμές κώδικα χρησιμοποιώντας MVC architectural components

Express.JS

Το Express.JS είναι ένα server-side web και mobile application frameworkτο οποίο χτίστηκε πάνω από το Node.JS server environment, παρέχοντας θεμελιώδη συστατικά για την υλοποίηση back-end web application functions σε JavaScript

Node.JS

Το Node.JS είναι ένα server environment για υλοποίηση scalable back-ends σε JavaScript

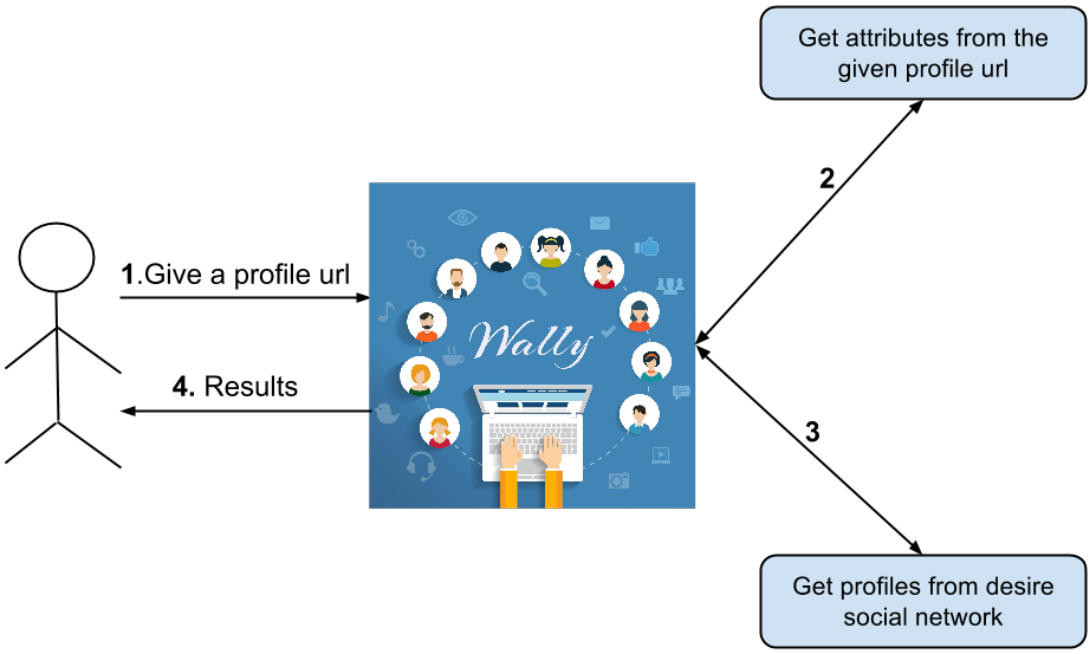
MongoDB

Η MongoDB είναι μια βάση δεδομένων η οποία χρησιμοποιά JSON-style documents για την αποθήκευση των δεδομένων. Μπορείς να χρησιμοποιήσεις την MongoDB με τον ίδιο τρόπο που θα έκανες και με την mySQL.

7.2.4 Manual

Όσο αφορά το Manual, ένας χρήστης μπορεί να δημιουργήσει λογαριασμό στον Wally βάζοντας μόνο το όνομα και ένα email

Στη συνέχεια το μόνο που χρειάζεται είναι να εισάγει ένα flickr profile url για να εξάγει αποτελέσματα στο facebook είτε να εισάγει ένα linkedIn profile url για να εξάγει αποτελέσματα για το twitter.



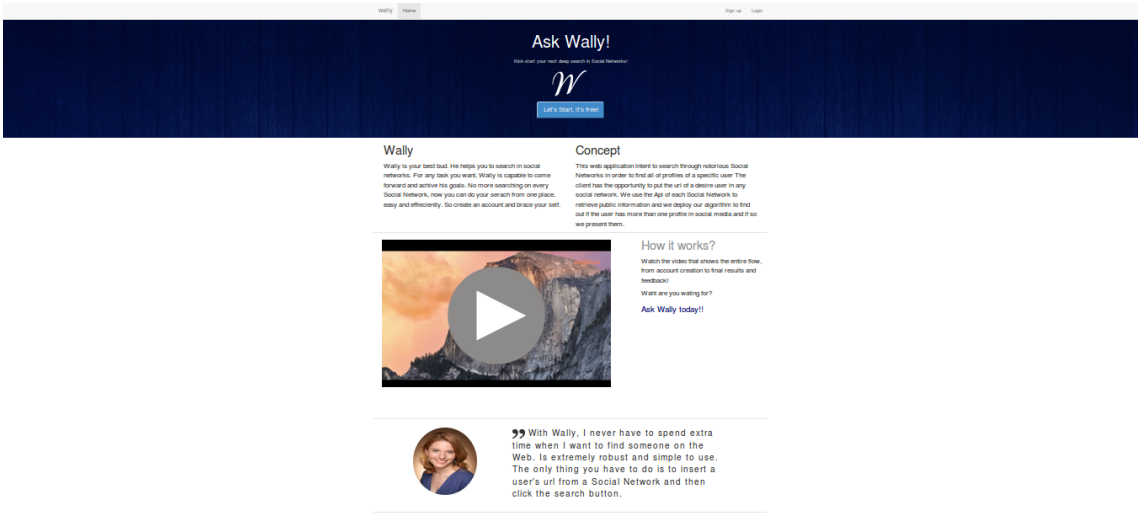
Σχήμα 7.4: Wally's Process

Πίο αναλυτικά στο σχήμα 7.4, παρουσιάζεται η ροή της εφαρμογής από την εισαγωγή κάποιου προφίλ που βάζει ο χρήστης μέχρι και την εξαγωγή των αποτελεσμάτων.

Σαν πρώτο βήμα είναι η εισαγωγή του προφίλ του χρήστη για τον οποίο προσπαθούμε να ανακαλύψουμε αν υπάρχουν προφίλ, του ίδιου χρήστη, στο επιθυμητό κοινωνικό δίκτυο. Στη συνέχεια η εφαρμογή μας βρίσκει και επιστρέφει στην εφαρμογή, όλα τα χαρακτηριστικά για τον χρήστη. Χρησιμοποιώντας αυτά τα χαρακτηριστικά αναζητάμε προφίλ στο επιθυμητό κοινωνικό δίκτυο και τα επιστρέφουμε. Μετέπειτα, εφαρμόζουμε την προσέγγισή μας και επιστρέφουμε στον χρήστη τα προφίλ μαζί με την πιθανότητα να ανήκουν στον χρήστη που εισάγαμε. Ο χρόνος περάτωσης τις εργασίας έχει ανώτατο όριο δέκα δευτερόλεπτα.

Στο σχήμα 7.5 παρουσιάζεται η κεντρική σελίδα της εφαρμογής όπου περιέχονται πληροφορίες για την εφαρμογή όπως ο σκοπός που δημιουργήθηκε και τις ευκολίες της οποίες προσφέρει. Στη συνέχεια στο σχήμα 7.6 παρουσιάζεται η διαδικασία εγγραφής καινούργιου χρήστη γεμίζοντας τα απαραίτητα πεδία, όπως όνομα ηλεκτρονική διεύθυνση και έναν κωδικό πρόσβασης.

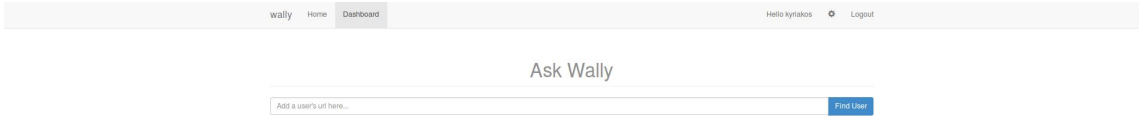
Έπειτα στο σχήμα 7.7 παρουσιάζεται η κύρια λειτουργία της εφαρμογής. Ο χρήστης έχει την δυνατότητα να εισάγει ένα url κάποιου χρήστη στο Flickr ή LinkedIn και να



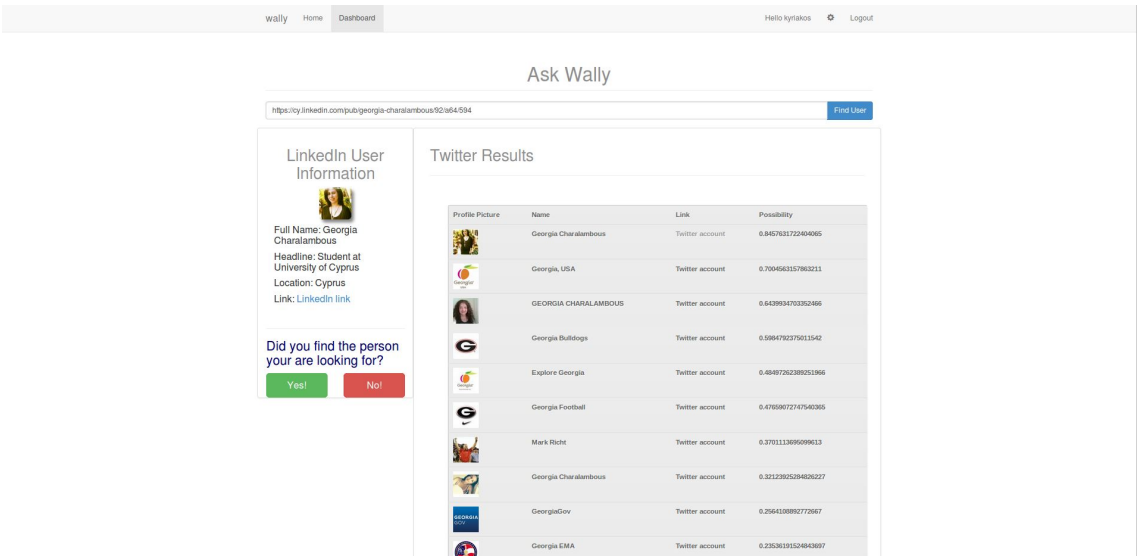
Σχήμα 7.5: Landing Page

του επιτρέψει σαν αποτέλεσμα τα προφίλ που ενδέχεται να ανήκουν στον ίδιο χρήστη για Facebook και Twitter αντίστοιχα, χρησιμοποιώντας την δική μας προσέγγιση. Τα επιστρεφόμενα προφίλ εμφανίζονται όπως φαίνεται και στο σχήμα 7.8

Σχήμα 7.6: Sing Up Page



Σχήμα 7.7: Search Page



Σχήμα 7.8: Results

Κεφάλαιο 8

Conclusion

8.1 Summary

Συνοψίζοντας όλη μας τη δουλεία, έχουμε συλλέξει δεδομένα για χρήστες κάνοντας χρήση ενός εργαλείου, το FriendFeed το οποίο επιτρέπει στους χρήστες να συναθροίζουν τα κοινωνικά τους δίκτυα. Έχοντας αυτά τα δεδομένα καταφέραμε να βρούμε τα κοινωνικά δίκτυα τα οποία χρησιμοποιούν οι χρήστες στο δείγμα μας.

Αφού αναγνωρίσαμε αυτά τα κοινωνικά δίκτυα, Facebook, Twitter, LinkedIn, Flickr, προχωρήσαμε στους συνδυασμούς που θα αναλύσουμε. Βασιζόμενοι στον αριθμό των χρηστών που έχουν κάποιο συνδυασμό, επιλέξαμε τους συνδυασμούς Facebook-Flickr and Twitter-LinkedIn.

Στη συνέχεια για κάθε έναν χρήστη που έχουμε, βρήκαμε όλα τα χαρακτηριστικά που μας παρέχονται από τα κοινωνικά δίκτυα όπως όνομα χρήστη, τοποθεσία, εικόνα προφίλ κ.α. Συνεπώς έχουμε τα προφίλ που ανήκουν στον ίδιο χρήστη. Μετέπειτα επεκτείναμε τα δεδομένα μας, βρίσκοντας προφίλ τα οποία δεν έχουν τον ίδιο χρήστη για να μπορέσουμε να αξιολογήσουμε τον αλγόριθμό μας.

Έπειτα, αναλύσαμε τα δεδομένα μας για να βρούμε ό,τι πληροφορία είχαμε για τους χρήστες για να μπορέσουμε να βρούμε τα χαρακτηριστικά που θα χρησιμοποιήσουμε στον αλγόριθμό μας. Αφού βρήκαμε τα χαρακτηριστικά εφαρμόσαμε την προσέγγιση μας με την logistic regression και μετά από ανάλυση των αποτελεσμάτων καταλήξαμε στο συμπέρασμα ότι έχουμε καλύτερα αποτελέσματα από ήδη υπάρχον μεθόδους που προσπαθούν να λύσουν το ίδιο πρόβλημα. Για Facebook-Flickr, έχουμε ποσοστό επιτυχίας 55% και Twitter-LinkedIn έχουμε ποσοστό επιτυχίας 82%

Τέλος υλοποιήσαμε ένα εργαλείο το οποίο προσφέρει στους χρήστες την δυνατότητα να βρίσκουν τα προφίλ του ίδιου χρήστη κάνοντας χρήση ενός profile url κάποιου κοινωνικού δικτύου . Επίσης παρουσιάσαμε το επιχειρηματικό πλάνο της εφαρμογής

καθώς και την αγορά/ζήτηση που υπάρχει και την μελλοντική εργασία καθώς βελτιώσεις που μπορεί να γίνουν.

8.2 Future Work

Σε αυτό το κεφάλαιο θα εισηγηθούμε κάποιες μελλοντικές βελτιώσεις όσο αφορά τον αλγόριθμο αλλά και την διαδικτυακή εφαρμογή.

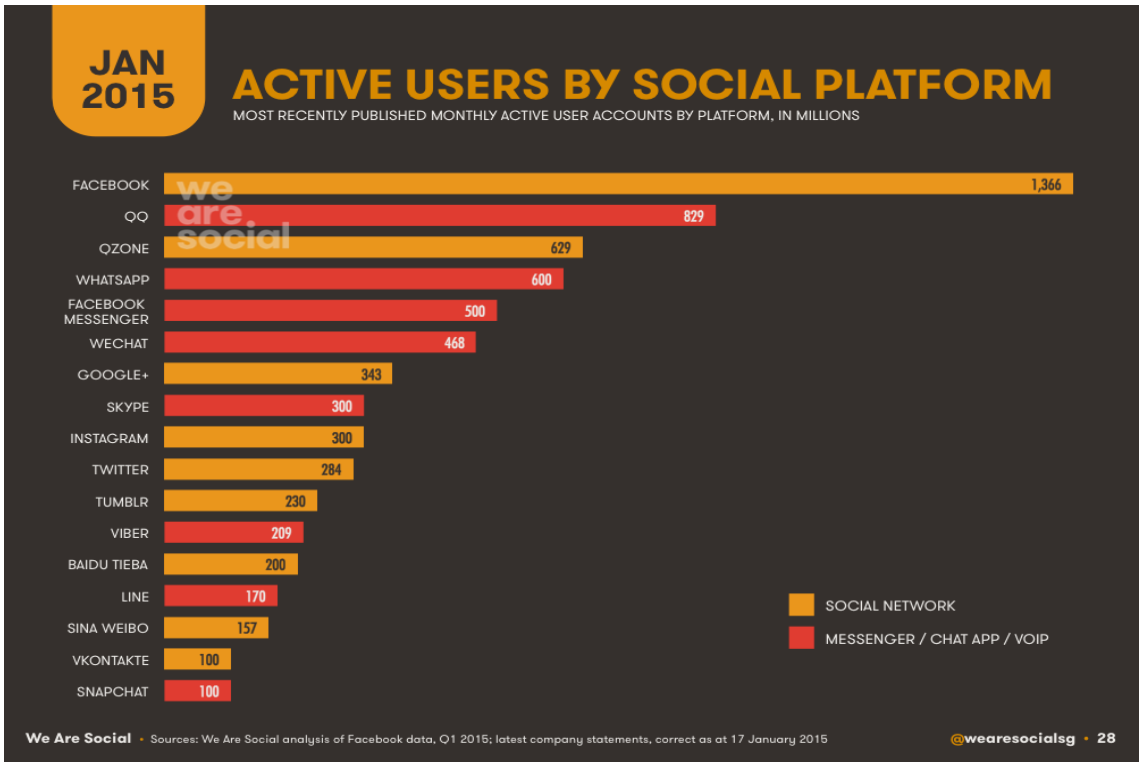
8.2.1 Optimizations

Όσο αφορά τον αλγόριθμό πρέπει να γίνει ανάλυση και στους υπόλοιπους συνδυασμούς μεταξύ των τεσσάρων κοινωνικών δικτύων που εισηγηθήκαμε. Και να παρουσιάζει αποτελέσματα και για τα τέσσερα κοινωνικά δίκτυα που αναλύσαμε.

Θα πρέπει να εισηγηθούμε εναλλακτικό εργαλείο για να αντικαταστήσει του FriendFeed.
Όπως:

1. Flavors.me
2. Mention.net
3. Rebel Mouse
4. Cyfe
5. Ming.ly
6. Postano
7. Stackla
8. Alternion
9. Jyst.us
10. HootSuite

Έπειτα πρέπει να εισάγουμε περισσότερα κοινωνικά δίκτυα για να κάνουμε τον αλγόριθμό μας πιο ελκυστικό προς τους χρήστες. Όπως παρουσιάζεται και στο σχήμα 8.1, όσο αφορά τα κοινωνικά δίκτυα, θα ήταν καλό να εισάγουμε το Goolge+ το Tumblr και το Instagram έτσι ώστε αν ένας χρήστης έχει ένα profile url οποιουδήποτε κοινωνικού δικτύου να παίρνει αποτελέσματα από όλα τα δημοφιλή κοινωνικά δίκτυα.



Σχήμα 8.1: Active users by Social Platform

Επίσης, σαν μελλοντική εργασία θα ήταν καλό να υλοποιηθούν οι προσεγγίσεις των άρθρων που αναφέραμε στο κεφάλαιο 2 και να τις εφαρμόσουμε στα δικά μας δεδομένα. Επειδή τα δεδομένα των εργασιών δεν είναι διαθέσιμα στο ευρύ κοινό.

Για την Διαδικτυακή Εφαρμογή θα πρέπει να επικεντρωθούμε στο HCI (Human Computer Interface) της εφαρμογής. Να εισάγουμε περισσότερα widgets που να παρέχουν στον χρήστη εύκολη και ευχάριστη πλοήγηση στην εφαρμογή. Και να προσπαθήσουμε να εκμεταλλευτούμε στο μέγιστο το MEAN Stack.

Bibliography

[1] E. Raad R. Chbeir and A. Dipanda. User profile matching in social networks. 5, page 297 304. In Proc. 13th International Conference on Network-Based Information Systems, 2010. Gifu Japan.

[2] R.Jain D.Sonnen. Social life networks. *IT Professional*, 13(5):8 11, 10 2011.

[3] Constantinos Fokianos. Logistic regression, 2009. [Logistic Regression Slides; Date: 21-Dec-2009].

[4] M.Motoyama G.Varghese. I seek you: Searching and matching individuals in social networks. *WIDM: Proceeding of the eleventh international workshop on Web information and data management.*, 2009.

[5] L. Ding L. Zhou T. Finin A. Joshi. How the semantic web is being used:an analysis of foaf. *Proceedings of the 38th International Conference on System Sciences*, 2005.

[6] Lynda. Mean stack, 2015. Build web apps with mean stack; Date: 02-May-2015].

[7] I. Polakis G. Kontaxis S. Antonatos E. Gessiou T. Petsas E. P. Markatos. Using social networks to harvest email addresses. *WPES : Proceedings of the 9th annual ACM Workshop on Privacy in the electronic society*, 2010.

[8] William B. Cavnar John M. Trenkle. N-gram-based text categorization. In *In Proceedings of SDAIR-94*, page 161 175. 3rd Annual Symposium on Document Analysis and Information Retrieva, 1994. Las Vegas.

[9] Wikipedia. Histogram matching, 2014. [Histogram matching; Date: 06-Feb-2014].

[10] Wikipedia. Damerau levenshtein distance, 2015. [Levenshtein distance; Date: 05-May-2015].

[11] Wikipedia. Decesion tree, 2015. [Decesion Tree; Date: 16-Apr-2015].

[12] Wikipedia. Digidal image correlation, 2015. [Digidal image correlation; Date: 27-Apr-2015].

[13] Wikipedia. Friendfeed aggregator, 2015. [FriendFeed Aggregator; Date: 16-Apr-2015].

[14] Wikipedia. Jaro distance, 2015. [Jaro Distance; Date: 17-Apr-2015].

[15] Wikipedia. Naive bayes, 2015. [Naive Bayes; Date: 16-Apr-2015].

[16] Wikipedia. Social nework, 2015. [Social Network; Date: 10-Apr-2015].

[17] Wikipedia. Web crawler, 2015. [Web Crawler; Date: 16-Apr-2015].

[18] G. w.You S.-w. Hwang Z. NieJ.-R. Wen. Socialsearch: enhancing entity search with social network matching. *EDBT/ICDT*, 2011.

Παράρτημα Α΄

R code

```
library(mlbench)
library(caret)
file = 'C:/Users/kyri/Desktop/SpringSemester-Thesis/FilesForR/TwitterLinkedIN/
Similarities3483_9000.csv'
read.csv(file) = ser
str(ser)
inTraining = createDataPartition(ser$class, p = .70, list = FALSE)
training = ser[ inTraining,]
testing = ser[-inTraining,]

Logistic regresion on training set
fit_ training = glm( class   X0+X1+X2+X3+X4+X5+X6 , family=binomial(logit),
data=training )
summary(fit_ training)
anova(fit_ training)
confint(fit_ training)
exp(coef(fit_ training))
exp(confint(fit_ training))
predict(fit_ training, type="response")
plot(predict(fit_ training, type="response"), xlab="Index", ylab="Predicted Val-
ues on training")
abline(h=0)
Logistic regresion on test set
fit_ testing= glm( class   X0+X1+X2+X3+X4+X5+X6 , family=binomial(logit),
data=testing)
```

summary(fit_ testing)
anova(fit_ testing)
confint(fit_ testing)
exp(coef(fit_ testing))
exp(confint(fit_ testing))
predict(fit_ testing, type="response")
plot(predict(fit_ testing, type="response"), xlab="Index", ylab="Predicted Values on testing")
abline(h=0)
Logistig regresion on all data set
fit= glm(class X0+X1+X2+X3+X4+X5+X6 , family=binomial(logit), data=ser)
summary(fit)
anova(fit)
confint(fit)
exp(coef(fit))
exp(confint(fit))
predict(fit, type="response")
plot(predict(fit, type="response"), xlab="Index", ylab="Predicted Values")
abline(h=0)
residuals(fit, type="deviance")
Decesion Tree
library(rpart)
read.csv(file) = ser
str(ser)
glm.out = glm(class name_ sim + name_ screen_ sim+summary_ descr_ sim + username_ screenname_ sim + headline_ descr_ sim+location_ sim+pic_ sim , family=binomial(logit), data=ser)
class X0+X1+X2+X3+X4+X5+X6
glm.out = glm(class X0+X1+X2+X3+X4+X5+X6, family=binomial(logit), data=ser)
fit = rpart(glm.out, data=ser, method='class')
printcp(fit)
plotcp(fit)
rsq.rpart(fit)
print(fit)

```
summary(fit)
plot(fit)
text(fit)
plot(fit, uniform=TRUE,main="Classification Tree for Profile Matching")
text(fit, all=TRUE, cex=.8)
pfit= prune(fit, cp=fit$cptable[which.min(fit$cptable[,"xerror"]),"CP"])
plot(pfit, uniform=TRUE,main="Pruned Classification Tree for Profile Matching")
text(pfit, all=TRUE, cex=.8)
feature Selection
library(caret)
read.csv(file) = ser
set.seed(999)
inTraining = createDataPartition(ser$class, p = .75, list = FALSE)
training = ser[ inTraining,]
testing = ser[-inTraining,]
fitControl = trainControl(method = "repeatedcv",number = 10,repeats = 10)
gbmFit1 = train(class   ., data = training,method = "gbm", trControl = fitControl,
verbose = FALSE)
trellis.par.set(caretTheme())
summary(gbmFit1)
plot(gbmFit1)
SVM
library(e1071)
library(klaR)
library(caret)
read.csv(file) = ser
ser= na.omit(ser)
inTraining = createDataPartition(ser$class, p = .70, list = FALSE)
training = ser[ inTraining,]
testing = ser[-inTraining,]
y = ser$class
x = subset(ser, select=-class)
modelAll = svm(x, y,type='C-classification')
predAll = predict(modelAll, x)
table(predAll , y)
print(modelAll )
```

```
plot( modelAll,ser)
yTrain = training$class
xTrain = subset(training , select =-class)
modelTrain = svm(xTrain , yTrain ,type='C-classification')
predTrain = predict(modelTrain , xTrain )
table(predTrain , yTrain )
yTest = testing$class
xTest = subset(testing , select =-class)
predTest = predict(modelTrain , xTest )
table(predTest , yTest )
Naive Bayes
library(e1071)
library(klaR)
library(caret)
read.csv(file) = ser
ser= na.omit(ser)
inTraining = createDataPartition(ser$class, p = .70, list = FALSE)
training = ser[ inTraining,]
testing = ser[-inTraining,]
model = naiveBayes(class ., data = training )
pred = predict(model,as.data.frame(testing))
table(pred,training[1:1,1])
print(pred)
```

Παράρτημα Β΄

Tools

Β΄.1 Εισαγωγή

Για την εκπόνηση της διπλωματικής εργασίας κάναμε χρήση από πολλές τεχνολογίες. Τόσο σε προγραμματιστικό επίπεδο όσο και σε ερευνητικό / στατιστικό επίπεδο. Επίσης σε πολλές περιπτώσεις χρειάστηκε να γίνει συνδυασμός αυτών των εργαλείων.

Β΄.2 Programming Tools

Η ανάπτυξη των εργαλείων έγινε σε λειτουργικό Linux και χρησιμοποιώντας το ολοκληρωμένο περιβάλλον ανάπτυξης Eclipse IDE . Όσο αφορά τα εργαλεία που χρησιμοποιήθηκαν που είχαν σχέση με προγραμματισμό είναι τα ακόλουθα:

Β΄.2.1 JAVA

Η κύρια υλοποίηση των αλγορίθμων και της μεθοδολογίας έγινε κάνοντας χρήση της γλώσσας προγραμματισμού JAVA. Η επιλογή έγινε λόγο του ότι τα προγράμματα που είναι γραμμένα σε JAVA τρέχουνε ακριβώς το ίδιο σε Windows, Linux, Unix and Macintosh. Επίσης λόγω του ότι είναι αντικειμενοστρεφείς και πολύ σταθερή.

Β΄.2.2 Python

Κατά τη διάρκεια κάποιων εργασιών για την ανάπτυξη της μεθοδολογίας, όπως της ανάλυσης των εικόνων προφίλ των χρηστών, χρειάστηκε να χρησιμοποιήσουμε την Python. Ο λόγος που χρησιμοποιήσαμε την Python ήταν επειδή έχει πολλές βιβλιοθήκες που είναι υλοποιημένες μόνο σε Python και παρέχουν πολλές δυνατότητες.

Συγκεκριμένα, στην περίπτωση των εικόνων μας παρείχε συναρτήσεις για να μπορέσουμε να πάρουμε πολλές πληροφορίες για μια εικόνα και επίσης συναρτήσεις που μας έδιναν την δυνατότητα να τροποποιήσουμε μια εικόνα.

Β′.2.3 HADOOP

Λόγω του ότι διαχειριστήκαμε μεγάλο όγκο δεδομένων, χρειαζόμασταν ένα εργαλείο για να μπορέσουμε να διαχειριστούμε εύκολα αυτά τα δεδομένα. Αυτό έγινε με το Hadoop το οποίο διαχειρίζεται με μεγάλη ευκολία μεγάλο όγκο δεδομένων χρησιμοποιώντας το Hadoop Distributed File System, το οποίο είναι ένα κατανεμημένο σύστημα αρχείων που αποθηκεύει κομμάτια δεδομένων σε διάφορες μηχανές και προσδίδει μεγάλο εύρος ζώνης και ταχύτητα. Ακολουθεί την λογική του Google MapReduce.

Β′.3 Statistical Tools

Για να μπορέσουμε να διαβάσουμε και να κατανοήσουμε τα αποτελέσματα μας, πρέπει να γίνει ανάλυση. Για αυτό τον σκοπό έγινε χρήση τον ακόλουθων εργαλείων:

Β′.3.1 R

Η R είναι μία γλώσσα προγραμματισμού που χρησιμοποιείται ευρέως για ανάπτυξη λογισμικών για στατιστική και επίσης για ανάλυση δεδομένων. Ένα άλλο σημαντικό στοιχείο του R είναι τα υψηλής ποιότητας γραφήματα και επίσης συμπεριλαμβανομένων των μαθηματικών συμβόλων.

Β′.4 Other Tools

Εκτός από τα προηγούμενα εργαλεία χρησιμοποιήθηκε και το ακόλουθο:

Β′.4.1 LaTex

Η συγγραφή της διπλωματικής, έγινε εξολοκλήρου στο LaTex. Το LaTex χρησιμοποιείται ευρέως στον ακαδημαϊκό χώρο, κυρίως λόγω της υψηλής ποιότητας στοιχειοθεσίας που παρέχει. Το τελικό αποτέλεσμα μπορεί να αποδοθεί σε μορφή pdf, dvi, ps κ.α. Το LaTex προσφέρει αυτοματοποίηση των περισσότερων πτυχών της στοιχειοθεσίας συμπεριλαμβανομένης της σελιδοποίησης, της βιβλιογραφίας, των περιεχομένων, αρίθμησης πινάκων, γραφικών παραστάσεων, εικόνων.