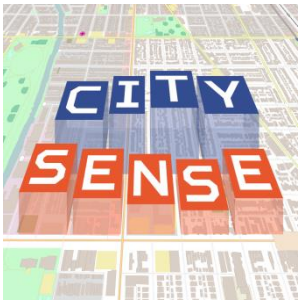
Πρόγραμμα "Σχέδιο Συμφωνίας Συμβιβασμού μεταξύ της Ελληνικής Δημοκρατίας και των εταιρειών Siemens AG και Siemens AE"

Έργο: Αλυσίδες Αξίας Δεδομένων σε Βιομηχανικά και Αστικά Περιβάλλοντα

Υποέργο: CitySense: Δυναμική, Διαδραστική και Πληθοποριστική Αστική Ανάλυση και Βιώσιμη Κινητικότητα



Ανάλυση Αστικών Δεδομένων: Μέθοδοι ανάλυσης και αξιοποίησης δεδομένων

Παραδοτέο Π2.1

Συγγραφείς

Χάρης Νάκος, Δημήτρης Παππάς, Δανάη Πλα-Καρύδη, Μαρία Ποντίκη, Γιάννης Σταύρακας, Χάρης Παπαγεωργίου, Μανώλης Τερροβίτης, Δημήτρης Τσιτσίγκος, Αλέξανδρος Εφεντάκης, Θοδωρής Δαλαμάγκας

Ημερομηνία: 23/12/2016

Επικαιροποιημένο την: 05/04/2017

«Αθηνά» - Ερευνητικό Κέντρο Καινοτομίας στις Τεχνολογίες της Πληροφορίας, των Επικοινωνιών και της Γνώσης

Πίνακας Περιεχομένων

1 Περίληψη

Τα κοινωνικά δίκτυα, τα διαθέσιμα ανοικτά δεδομένα, και οι μεγάλης κλίμακας διαδικτυακές προγραμματιστικές διεπαφές (massive online APIs) παρέχουν τεράστιες ποσότητες δεδομένων για την περιβάλλουσα τοποθεσία μας, ειδικά για πόλεις και αστικές περιοχές. Δυστυχώς, η πλειοψηφία των πρότερων εφαρμογών και ερευνητικών εγχειρημάτων εστίαζε συνήθως σε ένα είδος δεδομένων σε σχέση με κάποιο άλλο, παρουσιάζοντας, συνεπώς, μία μονομερή και ατελή οπτική για κάθε υπό μελέτη τοποθεσία, εξουδετερώνοντας έτσι τα προτερήματα τέτοιων προσεγγίσεων. Για την καταπολέμηση του προβλήματος αυτού, το παρόν έγγραφο παρουσιάζει το εργαλείο ανάπτυξης CitySense (CitySense framework). Το CitySense ταυτόχρονα συνδυάζει δεδομένα από διοικητικές πηγές (λόγου χάρη από δημόσιες υπηρεσίες), από μεγάλης κλίμακας προγραμματιστικές διεπαφές σημείων ενδιαφέροντος (Google Places, Foursquare), και από κοινωνικές υπηρεσίες δημοσίευσης περιεχομένου (Twitter), με σκοπό την ενιαία προβολή όλων των διαθέσιμων πληροφοριών για μία αστική περιοχή, σε μία διαισθητική και εύκολη σε χρήση πλατφόρμα διαδικτυακής εφαρμογής. Το Παραδοτέο Π2.1 θα παρουσιάσει τις προκλήσεις, από τεχνική και σχεδιαστική σκοπιά, της προσπάθειας αυτής, και το κατά πόσον αυτές οι διαφορετικές και αποκλίνουσες πηγές πληροφορίας μπορούν να συνδυαστούν, ώστε να παράσχουν μία ακριβή και πολύπλευρη οπτικοποίηση για την περίπτωση χρήσης που επιλέχθηκε, την αστική περιοχή του Chicago των Ηνωμένων Πολιτειών Αμερικής.

2 Εισαγωγή

Η εμφάνιση των κοινωνικών δικτύων, των πλατφορμών δημοσίευσης περιεχομένου, των εφαρμογών check-in και των κινητών τηλεφώνων / συσκευών GPS τα τελευταία χρόνια έχει δημιουργήσει τεράστιες ποσότητες δεδομένων σχετικές με την τοποθεσία των χρηστών. Για την εκμετάλλευση αυτών των ραγδαία αυξανόμενων δεδομένων, η πρόσφατη έρευνα έχει εστιάσει στη χρήση της γεωγραφικής πλευράς της πληροφορίας για την ανίχνευση γεγονότων, την ανάλυση συναισθήματος των χρηστών, την αποσαφήνιση τόπου-ονόματος, τον προσδιορισμό δημοφιλών σημείων ενδιαφέροντος και της χρονικής μεταβολής αυτών, τον προσδιορισμό και την οπτικοποίηση των τυπικών μοτίβων κίνησης των χρηστών κατά τη διάρκεια της ημέρας, όπως και τη βελτίωση των υπάρχοντων χαρτών πόλεων. Η γεωγραφική πληροφορία που προέρχεται από την εθελοντική συνεισφορά των διαδικτυακών χρηστών είναι, ωστόσο, ανακριβής από τη φύση της και για αυτό το λόγο θα πρέπει να χρησιμοποιείται με ιδιαίτερη προσοχή στο πλαίσιο κρίσιμων εφαρμογών.

Παρομοίως, η αυξανόμενη ανάγκη για αποδοτικές υπηρεσίες βασισμένες στην τοποθεσία και για αποτελεσματική διαδικτυακή διαφήμιση οδήγησε τους κυρίαρχους παρόχους υπηρεσιών στο διαδίκτυο (λόγου χάρη Google, Here, Bing, Foursquare) στην αποθήκευση και διάθεση πληροφοριών για σημεία ενδιαφέροντος (Points of Interest) στους χρήστες των υπηρεσιών αυτών, συνήθως μέσω της χρήσης διαδικτυακών προγραμματιστικών διεπαφών. Μία τέτοια προσέγγιση παρουσιάζει αρκετά προτερήματα, καθώς οι χρήστες δεν έχουν απλά πρόσβαση σε πληροφορία για τα κοντινά σε αυτούς σημεία ενδιαφέροντος, αλλά έχουν επίσης τη δυνατότητα να διαβάσουν και να παράσχουν κριτικές, ή να ενημερώσουν τους φίλους τους σχετικά με την τρέχουσα τοποθεσία τους. Οι ίδιες διαδικτυακές υπηρεσίες επιτρέπουν στους ιδιοκτήτες καταστημάτων και τις εταιρείες να διαφημίσουν τα καταστήματα και τις υπηρεσίες που παρέχουν. Όπως και σε κάθε εμπορικό προϊόν, ωστόσο, υπάρχουν περιορισμοί στη χρήση αυτών των προγραμματιστικών διεπαφών, με αποτέλεσμα να προσφέρεται στους χρήστες μία περιορισμένη σε τοπική κλίμακα οπτική της υπάρχουσας υποδομής της πόλης, η οποία δεν είναι δυνατόν να χρησιμοποιηθεί άμεσα για την εξαγωγή επιπρόσθετης πληροφορίας σε περιοχές ευρύτερης κλίμακας της πόλης.

Η κίνηση για ανοικτά δεδομένα, από μία άλλη πλευρά, υποστηρίζει ότι πρέπει οι πολίτες να έχουν πρόσβαση στα δεδομένα που συλλέγονται από κυβερνητικές υπηρεσίες, καθώς ουσιαστικά οι ίδιοι οι πολίτες χρηματοδοτούν τη συλλογή δεδομένων μέσω της φορολογίας. Ένα δεύτερο ισχυρό επιχείρημα είναι ότι η δημόσια πρόσβαση στα κυβερνητικά δεδομένα βοηθά τους ιδιώτες και τις εταιρείες να δημιουργούν εφαρμογές, οι οποίες τονώνουν την οικονομία, και να παρέχουν καλύτερες υπηρεσίες στους πολίτες, χωρίς επιπρόσθετο κόστος. Κάποιες χώρες και κάποιες πόλεις έχουν δημοσιεύσει τέτοια δεδομένα, τα οποία παρέχουν μία εναλλακτική οπτική των αστικών περιοχών. Παρόλο που αυτά τα ανοικτά δεδομένα είναι επίσημα, επιμελημένα, εξαιρετικής ποιότητας, και αδύνατον να συλλεχθούν από ιδιώτες, έχουν το προφανές μειονέκτημα ότι δεν μπορούν να είναι πραγματικού χρόνου, δεν είναι συνήθως προσβάσιμα μέσω προγραμματιστικών διεπαφών, και, το πιο σημαντικό, ενδέχεται να ενημερώνονται σε μη τακτά χρονικά διαστήματα (λόγου χάρη τα δημογραφικά δεδομένα), με αποτέλεσμα να υπάρχει ο κίνδυνος να είναι παρωχημένα.

Οι τρεις προαναφερθείσες πηγές πληροφορίας, δηλαδή η εθελοντικά διαθέσιμη γεωγραφική πληροφορία, τα διαδικτυακά δεδομένα σημείων ενδιαφέροντος, και τα επίσημα ανοικτά δεδομένα, έχουν, η καθεμία, τα προτερήματα και τις αδυναμίες τους, όσον αφορά την ακρίβεια των δεδομένων, το ρυθμό ενημέρωσης, την ευκολία χρήσης, και τη διαθεσιμότητα. Παρομοίως, οι εφαρμογές ή η έρευνα που κάνουν χρήση και βασίζονται σε έναν μόνον από αυτούς τους τύπους δεδομένων προσφέρουν μία μονομερή και ανακριβή οπτική της πραγματικότητας, η οποία ενδέχεται να είναι παραπλανητική. Για την

αντιμετώπιση αυτού του φαινομένου, παρουσιάζουμε το εργαλείο ανάπτυξης CitySense, το οποίο χρησιμοποιεί ανοικτά δεδομένα από διοικητικές πηγές, διαδικτυακές προγραμματιστικές διεπαφές σημείων ενδιαφέροντος, και κοινωνικές εφαρμογές δημοσίευσης περιεχομένου (tweets), με σκοπό την παροχή μίας ενιαίας οπτικής του παραδείγματος χρήσης που επιλέξαμε, της αστικής περιοχής του Chicago. Η αντίστοιχη διαδικτυακή εφαρμογή είναι διαθέσιμη στη διεύθυνση http://citysense.imis.athena-innovation.gr:8080/citysense/ και μπορεί να προβληθεί σε οποιονδήποτε σύγχρονο δικτυακό φυλλομετρητή (Chrome, Firefox). Δόθηκε έμφαση στο πώς μπορούν αποδοτικά να συναθροιστούν χωρικά, να οπτικοποιηθούν, και να παρουσιαστούν στον τελικό χρήστη, με καλαίσθητη και διαισθητική απεικόνιση, τα διαθέσιμα δεδομένα για καθεμία από αυτές τις τρεις πηγές, με ελάχιστη παρέμβαση στα δεδομένα, έτσι ώστε ο χρήστης να μπορεί ελεύθερα να ερμηνεύσει αυτή την πληροφορία κατά βούληση. Η εφαρμογή CitySense θα μπορούσε, έτσι, να επεκταθεί με επιπρόσθετη λειτουργικότητα με ελάχιστο κόπο. Το CitySense είναι, εν γένει, ένα δυναμικό σύστημα παρουσίασης αστικών περιοχών, το οποίο ενσωματώνει διάφορα σύνολα δεδομένων σχετιζόμενα με μία αστική περιοχή, παρέχοντας μία πλούσια οπτικοποίηση της ζωής μιας πόλης.

Ας θεωρήσουμε, για παράδειγμα, την περίπτωση ενός νεοφερμένου στην πόλη, ο οποίος πρέπει να αναζητήσει σπίτι σε μία άγνωστη περιοχή. Για τον εντοπισμό και αριθμητικό περιορισμό των γειτονιών προς αναζήτηση, θα πρέπει να απαντηθούν συγκεκριμένες ερωτήσεις. Αυτές οι ερωτήσεις ενδέχεται να συμπεριλαμβάνουν κριτήρια όπως είναι οι εκπαιδευτικές μονάδες ("Πού βρίσκονται οι πιο δημοφιλείς γειτονιές με κατοικίες οι οποίες διαθέτουν υψηλού επιπέδου εκπαιδευτικές μονάδες;") και η ασφάλεια ("Πού βρίσκεται η περιοχή στο κέντρο της πόλης η οποία διαθέτει τα λιγότερα μέτρα για την καταπολέμηση της εγκληματικότητας;"). Ως ένα άλλο παράδειγμα, ας θεωρήσουμε την περίπτωση ενός οργανωτή περιηγήσεων, οποίος επιθυμεί να παρακολουθεί την τουριστική δραστηριότητα σε μία πόλη, για να μπορεί να προσφέρει βελτιωμένα πακέτα και υπηρεσίες περιηγήσεων. Η παρακολούθηση, ωστόσο, της μεγάλης κλίμακας τουριστικής δραστηριότητας με τη χρήση παραδοσιακών μεθόδων θα απαιτούσε πολλή προσπάθεια, την εξέταση πληθώρας πηγών σε τακτά χρονικά διαστήματα, με μεγάλο κόστος και σπατάλη χρόνου για επιτόπου έρευνα.

3 Σχετικές Εφαρμογές

Τα τελευταία χρόνια, καθώς τα δεδομένα από συστήματα διαμοιρασμού τοποθεσίας αυξάνονται διαρκώς, οι ερευνητές έχουν προτείνει μία ευρεία ποικιλία μεθόδων για "αστική ανίχνευση", βασισμένων σε δεδομένα τοποθεσίας προερχόμενα από πηγές κάθε είδους: δημοσιεύσεις σε κοινωνικά δίκτυα και check-in, τηλεφωνική δραστηριότητα, δεδομένα κίνησης ταξί, δημογραφικά δεδομένα, κ.λπ. Οι επιστήμονες έχουν συνδυάσει κοινωνικές επιστήμες, την επιστήμη της πληροφορικής, και εργαλεία εξόρυξης δεδομένων, με σκοπό την εξαγωγή χρήσιμης γνώσης όσον αφορά τη ζωή των πόλεων. Ο Cranshaw [7] προσπάθησε να αποκαλύψει τις δυναμικές μιας πόλης βασιζόμενος στη δραστηριότητα των κοινωνικών δικτύων, ενώ στα [8, 9] οι συγγραφείς χαρακτήρισαν υποπεριοχές των πόλεων με την εξόρυξη σημαντικών μοτίβων που προέκυψαν από γεωχωρικά επισημειωμένα tweets. Οι Frias και Martinez [10] επικεντρώθηκαν στο συμπερασμό χρήσεων γης και σημείων ενδιαφέροντος σε μία συγκεκριμένη αστική περιοχή βασιζόμενοι σε μοτίβα δημοσίευσης tweets, και ο Noulas [11] ανέλυσε τις δυναμικές των check-in χρηστών, για την εξόρυξη χρήσιμων χωρικο-χρονικών μοτίβων με σκοπό την ανάλυση αστικών χώρων. Έχει γίνει πολλή δουλειά στο πεδίο της χρήσης κειμενικού και σημασιολογικού περιεχομένου από κοινωνικά δίκτυα με σκοπό την αστική ανάλυση. Ο Pozdnoukhov [12], για παράδειγμα, διεξήγαγε χωρική ανάλυση πραγματικού χρόνου του θεματικού περιεχομένου από ρεύματα tweets. Ο Noulas [13] πρότεινε, συν τοις άλλοις, τη σύγκριση των αστικών γειτονιών με τη χρήση σημασιολογικής πληροφορίας συνδεδεμένης με τα μέρη στα οποία οι άνθρωποι κάνουν check-in, ενώ ο Kling [14] εφάρμοσε ένα πιθανοτικό θεματικό μοντέλο με σκοπό τη λήψη της αποσύνθεσης του ρεύματος των ψηφιακών ιχνών σε ένα σύνολο αστικών θεμάτων σχετιζόμενων με τις ποικίλες δραστηριότητες των πολιτών με τη χρήση δεδομένων από το Foursquare και το Twitter. Οι Grabovitch και Zuyev [17] μελέτησαν τη συσχέτιση ανάμεσα στο κειμενικό περιεχόμενο και τις γεωχωρικές τοποθεσίες στα tweets και ο Kamath [16] χρησιμοποίησε τη χωρικο-χρονική διάδοση των hashtags για το χαρακτηρισμό τοποθεσιών. Μεθοδολογίες πρόβλεψης έχουν εκτενώς χρησιμοποιήσει γεωχωρικά επισημειωμένο κοινωνικό περιεχόμενο. Ο Kinsella [15], για παράδειγμα, δημιούργησε γλωσσικά μοντέλα των τοποθεσιών που εξήχθησαν από γεωχωρικά επισημειωμένα δεδομένα του Twitter, με σκοπό την πρόβλεψη της τοποθεσίας κάποιου συγκεκριμένου tweet, στα [18, 19, 20, και 22] οι συγγραφείς αποσκοπούσαν στη μοντελοποίηση της φιλίας ανάμεσα σε χρήστες αναλύοντας τα ίχνη των τοποθεσιών τους, και ο Cheng [21] εκτίμησε την τοποθεσία σε επίπεδο πόλης ενός χρήστη του Twitter βασισμένος αποκλειστικά στο περιεχόμενο των tweets του χρήστη. Οι ερευνητές έχουν, επίσης, εστιάσει στην ανίχνευση τάσεων και γεγονότων με την ανίχνευση συσχετίσεων ανάμεσα σε θέματα και τοποθεσίες [23, 24]. Έχουν δημοσιευτεί πολλές δουλειές τελευταία που εστιάζουν σε μοτίβα αστικής κινητικότητας. Ο Veloso [25], για παράδειγμα, ανέλυσε τις τροχιές απο δεδομένα κίνησης ταξί στη Λισαβόνα για την εξερεύνηση της σχέσης κατανομής ανάμεσα σε τοποθεσίες επιβίβασης και τοποθεσίες αποβίβασης. Στο [26] οι συγγραφείς εξερεύνησαν τις αναλυτικές μεθοδολογίες πραγματικού χρόνου για χωρικο-χρονικά δεδομένα καθημερινών μοτίβων κίνησης των πολιτών σε αστικό περιβάλλον και στα [27, 28, 29, 30, και 31] οι συγγραφείς χρησιμοποίησαν τα δεδομένα τροχιών κίνησης των χρηστών κινητών τηλεφώνων για τη μελέτη των δυναμικών πόλης και της ανθρώπινης κινητικότητας, ενώ στα [32, 33, 34, και 35] αναλύεται η ανθρώπινη κινητικότητα με τη χρήση δεδομένων κοινωνικών δικτύων. Ένα άλλο πεδίο που συνδέεται με την αστική ανάλυση είναι οι γεωδημογραφικές κατηγοριοποιήσεις, οι οποίες αναπαριστούν κατηγοριοποιήσεις μικρών περιοχών που παρέχουν συνοπτικούς δείκτες των κοινωνικών, οικονομικών, και δημογραφικών χαρακτηριστικών των γειτονιών [36]. Στην περιοχή των δημογραφικών τοποθεσιών και της κοινωνικο-οικονομικής πρόβλεψης και συσχέτισης, οι

ερευνητές έχουν προτείνει μία ποικιλία μεθόδων βασιζόμενων σε γεωχωρικά επισημειωμένα δεδομένα κοινωνικών δικτύων [37, 38, και 39].

Ως αποτέλεσμα, έχει αναπτυχθεί ως τώρα μία ευρεία ποικιλία εφαρμογών που περιγράφουν τη ζωή αστικών περιοχών. Το EvenTweet [40], για παράδειγμα, είναι ένα σύστημα για την ανίχνευση εντοπισμένων χωρικά γεγονότων σε πραγματικό χρόνο από ένα ρεύμα δεδομένων του Twitter και για την ανίχνευση της εξέλιξης των γενονότων αυτών με την πάροδο του χρόνου. Το "One million Tweet Map" [41] είναι, επίσης, μία διαδικτυακή εφαρμογή που προβάλλει τα τελευταία ένα εκατομμύριο tweets πάνω στον παγκόσμιο χάρτη σε πραγματικό χρόνο. Ο χάρτης ανανεώνεται κάθε δευτερόλεπτο, πετώντας τα παλιότερα είκοσι tweets, και σχεδιάζοντας τα πιο πρόσφατα είκοσι tweets, κρατώντας το συνολικό αριθμό tweets γύρω στο ένα εκατομμύριο, δείχνοντας τα tweets ομαδοποιημένα σε περιοχές ανά τον κόσμο, ενώ οι χρήστες μπορούν να κάνουν ζουμ μέσα και έξω στο χάρτη, προκαλώντας τον επανυπολογισμό των ομαδοποιήσεων. Η εφαρμογή "tweepsmap" [42], με τη σειρά της, παρέχει στους χρήστες αποδοτικά στοιχεία ανάλυσης και τη δυνατότητα διαχείρισης για γεωγραφικά δεδομένα του Twitter, ενώ το "trendsmap" [43] και το "tweetmap" [44] παρουσιάζουν τις γεωγραφικά εντοπισμένες τελευταίες τάσεις στο Twitter πάνω σε χάρτη. Στο πεδίο της γραφικής οπτικοποίησης Δημογραφικών Στοιχείων Απογραφής Αστικών Περιοχών, το "Mapping America: Every City, Every Block" [45] δίνει τη δυνατότητα στους χρήστες να εξερευνήσουν τοπικά δεδομένα από την Επισκόπηση της Αμερικανικής Κοινότητας του Γραφείου Απογραφών, η οποία βασίζεται σε στοιχεία από το 2005 έως το 2009. Το "Social Explorer" [46], τέλος, παρέχει εργαλεία βασισμένα σε χάρτη για την γραφική εξερεύνηση δημογραφικής πληροφορίας, η οποία συμπεριλαμβάνει την αρχή Απογραφής Η.Π.Α., την αρχή Επισκόπησης Αμερικανικής Κοινότητας, την αρχή Απογραφής Ηνωμένου Βασιλείου, την αρχή Καναδικής Απογραφής, την Ευρωπαϊκή Στατιστική Αρχή, την αρχή "Uniformed Crime Report" του Ομοσπονδιακού Γραφείου Ερευνών, τα αποτελέσματα αμερικανικών εκλογών, την αρχή "Religious Congregation Membership Study", την αρχή "World Development Indicators".

Παρόλο που οι προαναφερθείσες δουλειές παρέχουν εις βάθος ματιά σε κάποιες πλευρές της ζωής σε μία αστική περιοχή, αδυνατούν να παράσχουν μία ενιαία και πιο γενική οπτική της πόλης και να δώσουν στο χρήστη τη δυνατότητα να απαντήσει ερωτήσεις διαδραστικά, συνδυάζοντας σύνολα δεδομένων. Το CitySense αποσκοπεί στην κάλυψη αυτών των κενών με την ενσωμάτωση πολλαπλών πηγών δεδομένων και με την παροχή μίας διαδραστικής διεπαφής χρήστη που να υποστηρίζει φίλτρα, επιλογές πολλαπλών οπτικών, και δυνατότητες προβολής πληροφορίας σε διάφορα επίπεδα λεπτομέρειας.

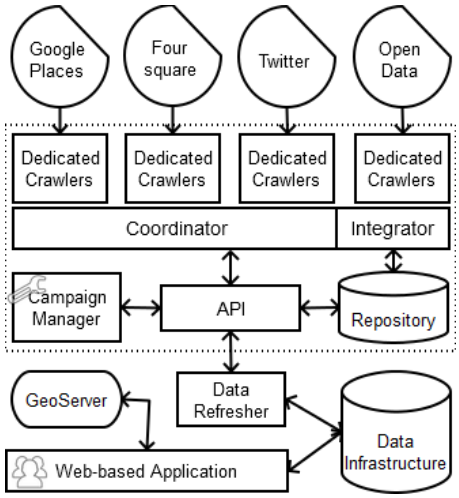
4 Αρχιτεκτονική του Συστήματος

Το CitySense είναι ένα δυναμικό σύστημα παρουσίασης αστικών περιοχών, το οποίο ενσωματώνει διάφορα σύνολα δεδομένων που σχετίζονται με μία αστική περιοχή και παρέχει μία πλούσια οπτικοποίηση της ζωής μιας πόλης. Η εφαρμογή μπορεί να απαντήσει ερωτήσεις σε πολλά επίπεδα, εκμεταλλευόμενη την ποικιλία των συνόλων δεδομένων που αναφέρονται σε μία πόλη και συνδυάζοντας ανομοιογενείς πηγές δεδομένων με εύκολο τρόπο. Οι χρήστες μπορούν να προβάλουν διάφορες πλευρές της ζωής της πόλης, στατικά ή στην πάροδο του χρόνου, για ολόκληρη την πόλη ή για κάθε τμήμα της, αναμειγνύοντας πηγές δεδομένων, με σκοπό την αποκάλυψη μοτίβων και πληροφορίας που δεν θα ήταν προφανή από την απλή παρατήρηση των συνόλων δεδομένων.

Η εφαρμογή CitySense στοχεύει στην παροχή ενός γρήγορου και εύκολου τρόπου για:

- το συνδυασμό ετερογενών πηγών δεδομένων σχετιζόμενων με διάφορες πλευρές της πόλης,
- το φιλτράρισμα δεδομένων και τον έλεγχο του επιπέδου λεπτομέρειας μέσω ενός περιβάλλοντος οπτικοποίησης βασισμένου σε χάρτη, και
- την απάντηση σε ερωτήσεις, την εξερεύνηση, και την ανακάλυψη χρήσιμης πληροφορίας με σκοπό την απόδοση της αίσθησης της πόλης.

Η αρχιτεκτονική του συστήματος παρουσιάζεται στην Εικόνα 1 και περιλαμβάνει τη διαδικτυακή εφαρμογή του CitySense (CitySense Web-based Application), την υποδομή δεδομένων (Data Infrastructure) και τις μονάδες ανανέωσης δεδομένων (Data Refresher), τον GeoServer, ο οποίος αναλύεται σε επόμενη ενότητα, και το υποσύστημα CityProfiler (το κουτί με διακεκομμένες γραμμές στο σχήμα), το οποίο υλοποιήθηκε για τη συλλογή των δεδομένων που σχετίζονται με την πόλη από τις πηγές δεδομένων, και το οποίο παρουσιάζεται λεπτομερώς παρακάτω.



Εικόνα 1. Αρχιτεκτονική του CitySense

Για την πιλοτική εφαρμογή επιλέχθηκε η πόλη του Chicago, λόγω της ποσότητας επίσημων δημογραφικών δεδομένων που είναι διαθέσιμα. Ανεξάρτητα από τα δεδομένα αυτά, ωστόσο, οι κάτοικοι του Chicago παρουσιάζουν έντονη δραστηριότητα και στα κοινωνικά δίκτυα, ενώ επαρκής αριθμός σημείων ενδιαφέροντος είναι, επίσης, διαθέσιμος.

Ο CityProfiler (περικλείεται από το κουτί με διακεκομμένες γραμμές στην Εικόνα 1) είναι ένα υποσύστημα του CitySense, υπεύθυνο για τη συλλογή δεδομένων που σχετίζονται με μία αστική περιοχή, από ετερογενείς πηγές. Η βασική του λειτουργία είναι η συλλογή όλων

των διαθέσιμων σημείων ενδιαφέροντος, των tweets, και των δημογραφικών δεδομένων που προέρχονται από τις υπηρεσίες της πόλης, και η αποθήκευση της πληροφορίας αυτής σε ένα αποθετήριο (repository) μαζί με τα σχετικά μετα-δεδομένα.

Ο CityProfiler παρέχει μία προγραμματιστική διεπαφή και μία γραφική διεπαφή, μέσω των οποίων μπορούν, αντίστοιχα, εφαρμογές και χρήστες να ορίσουν και να εκτελέσουν νέες καμπάνιες συλλογής δεδομένων. Κάθε καμπάνια, η οποία ορίζεται από συγκεκριμένες παραμέτρους, αναπαριστά μία ανεξάρτητη συλλογή δεδομένων. Αυτές οι παράμετροι ελέγχουν τους επιμέρους μηχανισμούς συλλογής, οι οποίοι συλλέγουν δεδομένα μέσω των διαθέσιμων προγραμματιστικών διεπαφών, και οι οποίες είναι οι εξής:

- Διάρκεια συλλογής: καθορίζει τη διάρκεια της καμπάνιας.
- Επιλογή μηχανισμού συλλογής: καθορίζει ποιοι από τους διαθέσιμους μηχανισμούς συλλογής (οι οποίοι αντιστοιχίζονται σε διακριτές πηγές δεδομένων) θα συμμετάσχουν στην καμπάνια.
- Τοποθεσία συλλογής: καθορίζει την τοποθεσία στην οποία θα γίνει η συλλογή.
- Επιλογή κατηγοριών: καθορίζει ποιες κατηγορίες σημείων ενδιαφέροντος μπορούν να εμφανιστούν, και επίσης λέξεις-κλειδιά στις οποίες θα βασιστεί η συλλογή. Οι λέξεις-κλειδιά είναι χρήσιμες στην περίπτωση που πρέπει να συλλεχθούν δεδομένα ξεπερνώντας γλωσσικές και πολιτιστικές παραλλαγές. Η επιλογή κατηγοριών δίνει, επίσης, τη δυνατότητα συλλογής όλων των σημείων ενδιαφέροντος μιας τοποθεσίας, ανεξάρτητα από την κατηγορία τους.
- Επιλογή συχνότητας συλλογής: κάποια από τα δεδομένα που έχουν συλλεχθεί χρειάζονται συστηματική ενημέρωση, εξαιτίας αλλαγών που ενδέχεται να συμβούν στα δημογραφικά δεδομένα και στα σημεία ενδιαφέροντος (μία καφετέρια μπορεί, λόγου χάρη, να μετατραπεί σε μπαρ, ή νέα σημεία ενδιαφέροντος μπορούν να εμφανιστούν). Το CitySense μπορεί να εκτελέσει επαναληπτικές καμπάνιες με μεγάλη διάρκεια, στις οποίες μπορούν να εκτελεστούν πολλαπλές συλλογές δεδομένων με τις ίδιες παραμέτρους. Η επιλογή συχνότητας καθορίζει, συνεπώς, πόσο συχνά πρέπει να γίνεται αυτόματη επανεκκίνηση της καμπάνιας.

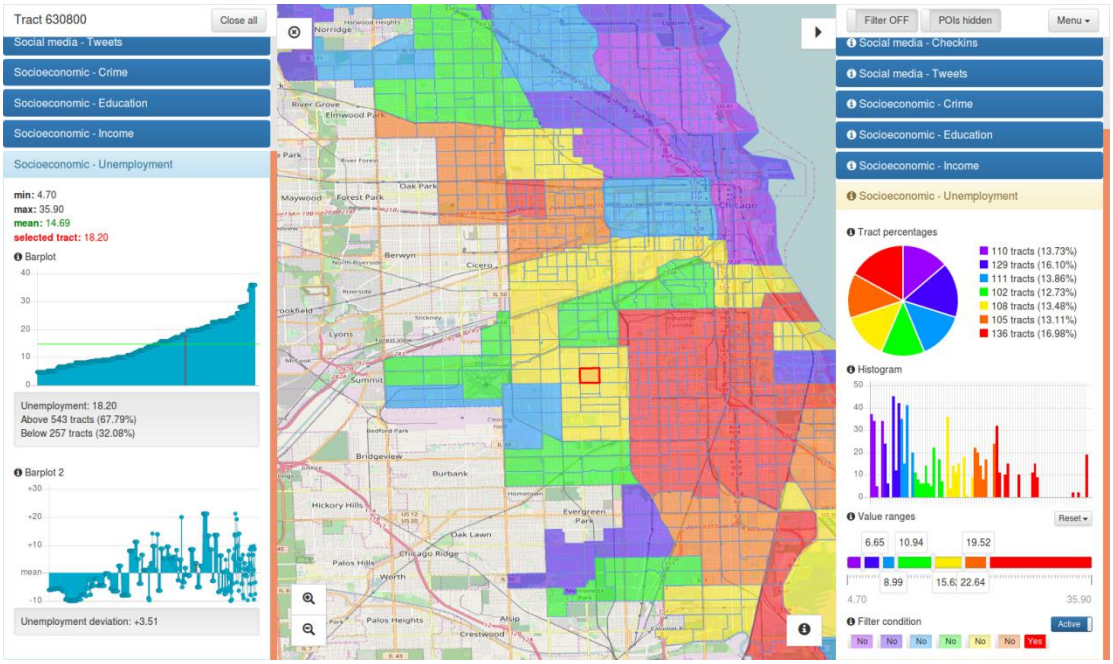
Ο CityProfiler μπορεί να εκτελεί παράλληλα πολλαπλές καμπάνιες, οπότε υπάρχει ανάγκη για ένα συντονιστή (Coordinator), όπως φαίνεται και στην Εικόνα 1, για τον έλεγχο των μηχανισμών συλλογής και τη διαχείριση κάθε καμπάνιας. Ο CityProfiler διαχειρίζεται, επίσης, τους πόρους με έξυπνο τρόπο, διασφαλίζοντας ότι ικανοποιούνται όλοι οι περιορισμοί που τίθενται από τις πηγές (λόγου χάρη σχετικά με τον μέγιστο αριθμό αιτημάτων ανά συγκεκριμένη χρονική περίοδο), και ότι αποφεύγονται επικαλυπτόμενα αιτήματα. Τα δεδομένα που έχουν συλλεχθεί καθαρίζονται για τον αποκλεισμό διπλοτύπων, και αποθηκεύονται προσωρινά σε ένα αποθετήριο.

Ο CityProfiler παρέχει ειδικό χειρισμό για ανοικτά δημογραφικά δεδομένα και για δεδομένα κοινωνικών δικτύων. Τα ανοικτά δημογραφικά δεδομένα δημοσιεύονται από κάποια δημόσια υπηρεσία σε διάφορες μορφές αρχείων (CSV, χωρισμένα με tabs, κ.λπ.). Ο CityProfiler, συνεπώς, τα συλλέγει και τα αποθηκεύει ως αρχεία σε ένα κατάλογο μαζί με μετα-πληροφορίες για την ημερομηνία και την πηγή συλλογής. Η συλλογή και αποθήκευση των δεδομένων κοινωνικών δικτύων γίνεται δυναμικά σε πραγματικό χρόνο, χρησιμοποιώντας ένα μηχανισμό συλλογής που κρατάει συνεχώς ενημερωμένο το περιεχόμενο που αντιστοιχίζεται σε ένα καθορισμένο κυλιόμενο παράθυρο μέσα στο χρόνο.

Στην επόμενη ενότητα παρουσιάζεται ο σχεδιασμός και η λειτουργία του συστήματος CitySense.

5 Σχεδιασμός και Λειτουργία του CitySense

Στην Εικόνα 2 παρουσιάζεται η διαδικτυακή διεπαφή χρήστη του CitySense. Το κεντρικό στοιχείο της οπτικοποίησης είναι ο χάρτης του Chicago. Ο χάρτης του Chicago χωρίζεται σε μικρότερα τμήματα, τα οποία ονομάζονται tracts. Τα tracts είναι υπάρχουσες διοικητικές υποδιαιρέσεις που χρησιμοποιούνται ήδη από τις κυβερνητικές υπηρεσίες της πόλης του Chicago. Το Chicago περιέχει 801 tracts, καθένα από τα οποία αντιπροσωπεύει μία μικρή περιοχή που θεωρείται ότι είναι σχετικά ομοιόμορφη και αντιστοιχίζεται ιδανικά σε περίπου 1200 νοικοκυριά (2000 – 4000 κατοίκους). Τα όρια των tracts επισημαίνονται με γαλάζια γραμμή και είναι πάντα ορατά στο χάρτη, ενώ όταν επιλέγεται ένα συγκεκριμένο tract τα όριά του επισημαίνονται με κόκκινη γραμμή.



Εικόνα 2. Διαδικτυακή διεπαφή χρήστη του CitySense

Αριστερά και δεξιά του χάρτη, το CitySense διαθέτει δύο μπάρες, οι οποίες παρέχουν δύο συμπληρωματικές όψεις του Chicago. Η πρώτη όψη, η οποία περιέχεται στη δεξιά μπάρα, παρέχει πληροφορίες και λειτουργίες σχετικές με την πόλη συνολικά. Χρησιμοποιώντας την όψη αυτή, οι χρήστες μπορούν να καθορίσουν επιλογές οπτικοποίησης και φιλτραρίσματος και να παρατηρήσουν τα αποτελέσματα τόσο στο χρωματισμό του χάρτη της πόλης όσο και στα διαγράμματα κατανομών που περιλαμβάνονται στην όψη. Η δεύτερη όψη, η οποία περιέχεται στην αριστερή μπάρα, παρέχει πληροφορίες και διαγράμματα σχετιζόμενα με το επιλεγμένο tract, το οποίο, όπως αναφέραμε, επισημαίνεται με κόκκινο περίγραμμα στο χάρτη. Η συγκεκριμένη όψη εμφανίζεται μόνο όταν επιλέγεται κάποιο tract και βοηθάει τους χρήστες να εντοπίσουν τα ειδικά χαρακτηριστικά κάθε tract και να το συγκρίνουν με τη γενικότερη εικόνα της πόλης. Οι δύο όψεις μπορούν να είναι ενεργές ταυτόχρονα, δίνοντας τη δυνατότητα στους χρήστες να παρατηρήσουν τα διαφορετικά σύνολα δεδομένων που περιλαμβάνονται και στις δύο όψεις, σε γενικότερο επίπεδο και σε επίπεδο tract την ίδια στιγμή.

Και οι δύο όψεις παρέχουν οπτικοποιήσεις και διαγράμματα σύμφωνα με τις προδιαγραφές του υπό εξέταση συνόλου δεδομένων. Για παράδειγμα, όπως φαίνεται στην Εικόνα 2, ο χρωματισμός του χάρτη και τα διαγράμματα οπτικοποιούν το σύνολο δεδομένων ανεργίας ("Socioeconomic – Unemployment"). Για την επιλογή του συνόλου

δεδομένων, ο χρήστης καλείται να επιλέξει ένα συρτάρι δεδομένων (data drawer). Τα συρτάρια δεδομένων για όλα τα σύνολα δεδομένων είναι προσβάσιμα και στις δύο όψεις ανεξάρτητα. Το καθένα από αυτά επισημαίνεται από τον τίτλο του συνόλου δεδομένων που εκπροσωπεί, μέσα σε σκούρο μπλε ορθογώνιο παραλληλόγραμμο. Τα διαθέσιμα σύνολα δεδομένων, και για τις δύο όψεις, είναι τα εξής:

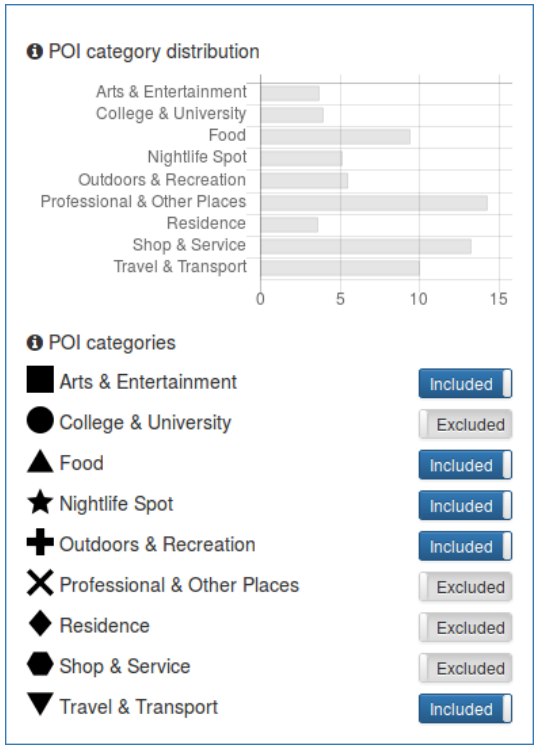
- Points of interest
- Health – Births
- Health – Infant mortality
- Social media – Checkins
- Social media – Tweets
- Socioeconomic – Crime
- Socioeconomic – Education
- Socioeconomic – Income
- Socioeconomic – Unemployment

Λόγω της διαφορετικής φύσης των συνόλων δεδομένων και, ταυτόχρονα, της ανάγκης για την κατά το δυνατόν ενιαία αντιμετώπιση όλων των συνόλων δεδομένων, όσο αφορά στο χρωματισμό του χάρτη αλλά και στα διαγράμματα τόσο στη δεξιά όσο και στην αριστερή μπάρα, δημιουργήθηκαν για κάποια σύνολα δεδομένων εξειδικεύμενοι επιλογείς δεδομένων. Κάθε επιλογέας δεδομένων, για ένα συγκεκριμένο σύνολο δεδομένων, εμφανίζεται στην κορυφή του συρταριού του συγκεκριμένου συνόλου δεδομένων, τόσο στη δεξιά όσο και στην αριστερή μπάρα. Το υπόλοιπο τμήμα του συρταριού, κάτω από τον επιλογέα δεδομένων (αν υπάρχει), είναι για όλα τα σύνολα δεδομένων το ίδιο. Αυτό ισχύει για κάθε μπάρα ξεχωριστά. Ας αναλύσουμε, λόγου χάρη, το σύνολο δεδομένων ανεργίας. Για την ανεργία είχαμε διαθέσιμο, για κάθε tract, το ποσοστό του εργατικού δυναμικού που είναι άνεργοι. Για την ανεργία δεν υπήρχε καμία παράμετρος που να δίνει τη δυνατότητα για παραπάνω ανάλυση του συνόλου δεδομένων. Δεν υπήρχε, για παράδειγμα, καμία χρονική παράμετρος, καμία παράμετρος φύλου, κ.λπ. Τα σύνολα δεδομένων τέτοιας φύσεως, που αντιστοίχιζαν, δηλαδή, απευθείας κάθε tract στην τιμή του μεγέθους για το συγκεκριμένο tract, δεν απαιτούσαν επιλογέα δεδομένων. Τέτοια σύνολα δεδομένων ήταν τα εξής:

- Health – Births
- Health – Infant mortality
- Socioeconomic – Education
- Socioeconomic – Income
- Socioeconomic – Unemployment

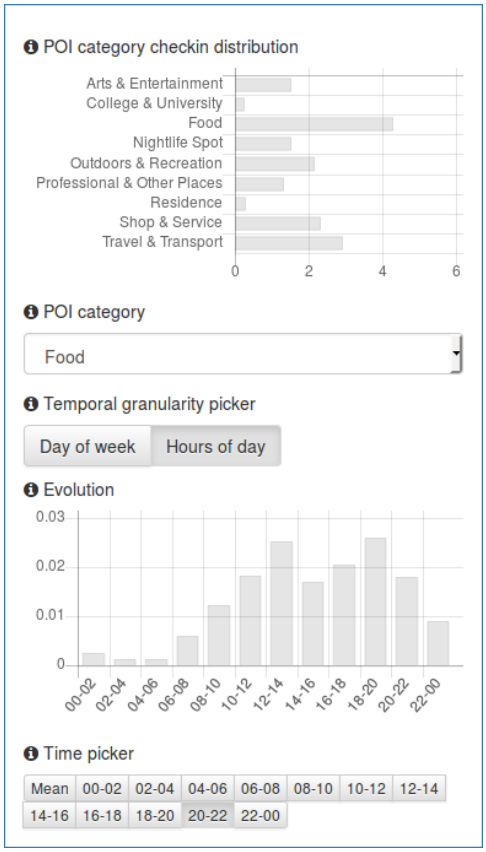
Τα παραπάνω σύνολα δεδομένων μπορούσαν να οπτικοποιηθούν απευθείας. Τα υπόλοιπα σύνολα δεδομένων απαιτούσαν, λόγω της ύπαρξης επιπρόσθετης πληροφορίας, κατηγορικής ή/και χρονικής, έναν εξειδικευμένο επιλογέα δεδομένων για το καθένα.

Για το σύνολο δεδομένων "Points of interest" ως επιλογέας δεδομένων χρησιμοποιήθηκε ένας κατηγορικός επιλογέας πολλαπλής επιλογής, για τον καθορισμό των κατηγοριών των σημείων ενδιαφέροντος που θα συμπεριληφθούν στην άθροιση για όλα τα tracts, ο οποίος παρουσιάζεται στην Εικόνα 3. Οι "Included" κατηγορίες σημείων ενδιαφέροντος θα συμπεριληφθούν στην άθροιση, ενώ οι "Excluded" κατηγορίες σημείων ενδιαφέροντος δεν θα συμπεριληφθούν.



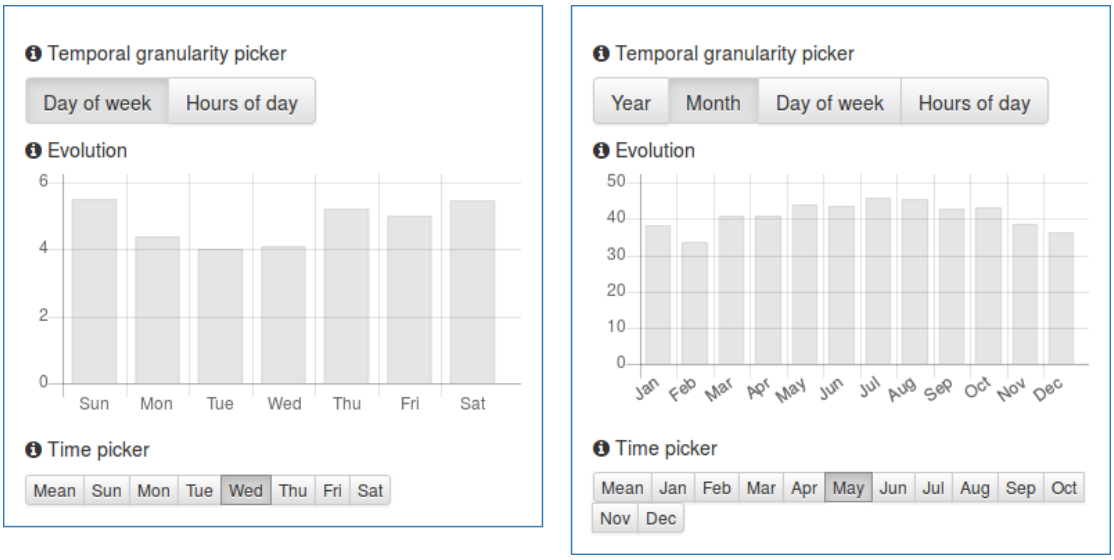
Εικόνα 3. Επιλογέας (κατηγορικός πολλαπλής επιλογής) δεδομένων PoIs

Για το σύνολο δεδομένων "Social media – Checkins", ως επιλογέας δεδομένων χρησιμοποιήθηκε ο συνδυασμός ενός κατηγορικού επιλογέα απλής επιλογής, για την επιλογή μίας κατηγορίας σημείων ενδιαφέροντος στην οποία έγιναν check-ins, και ενός χρονικού επιλογέα, για την επιλογή της ημέρας της εβδομάδας ή των ωρών της ημέρας για την ομαδοποίηση των check-ins, όπως παρουσιάζεται στην Εικόνα 4.



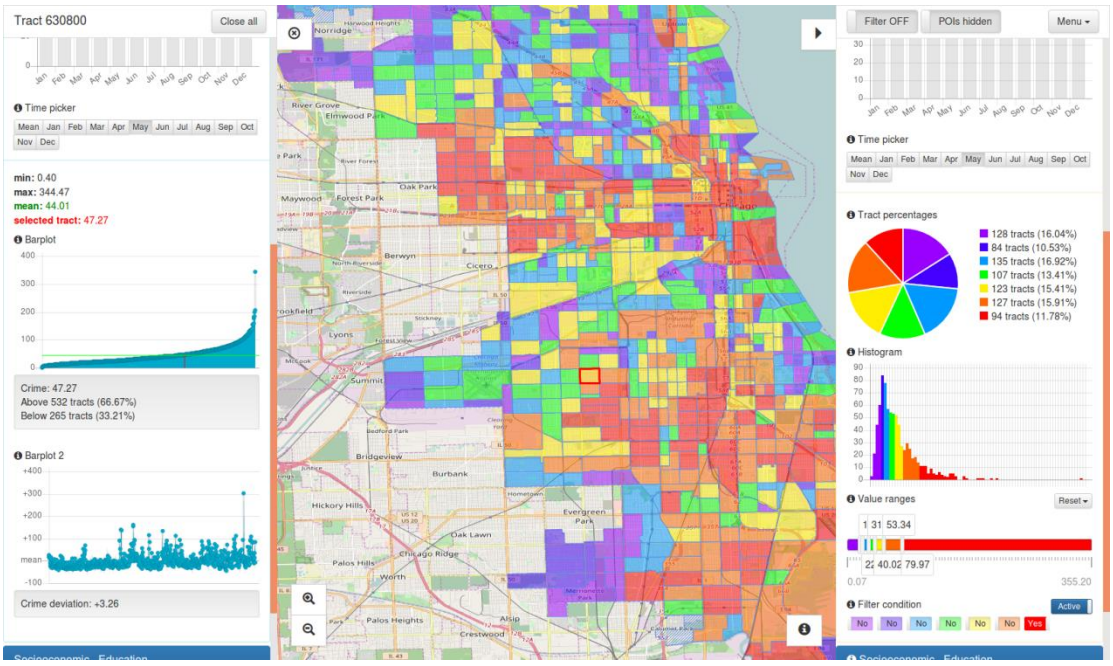
Εικόνα 4. Επιλογέας (κατηγορικός απλής επιλογής και χρονικός) δεδομένων Check-ins

Για τα σύνολα δεδομένων "Social media – Tweets" και "Socioeconomic – Crime", ως επιλογείς δεδομένων χρησιμοποιήθηκαν χρονικοί επιλογείς, όπως φαίνεται στην Εικόνα 5. Για την ομαδοποίηση των tweets υπάρχει η δυνατότητα επιλογής της ημέρας της εβδομάδας ή των ωρών της ημέρας, ενώ για την ομαδοποίηση των εγκλημάτων υπάρχει η δυνατότητα επιλογής του έτους, του μήνα, της ημέρας της εβδομάδας, ή των ωρών της ημέρας.



Εικόνα 5. Επιλογείς (χρονικοί) δεδομένων Tweets (αριστερά) και Crimes (δεξιά)

Το αποτέλεσμα της εφαρμογής του επιλογέα δεδομένων, αν υπάρχει για κάποιο μέγεθος, είναι η αντιστοίχιση κάθε tract, για όλα τα tracts, στην τιμή του μεγέθους για το συγκεκριμένο tract, με βάση τη συνθήκη του επιλογέα δεδομένων. Αυτή η αντιστοίχιση μπορεί να απεικονιστεί πλέον σε γραφήματα, τόσο στην αριστερή όσο και στη δεξιά μπάρα, και στο χρωματισμό του χάρτη. Στην Εικόνα 2 παρουσιάζονται τα γραφήματα και ο χρωματισμός του χάρτη με βάση το ποσοστό ανεργίας. Στην Εικόνα 6 παρουσιάζονται τα γραφήματα και ο χρωματισμός του χάρτη με βάση τα εγκλήματα που έγιναν μήνα Μάιο (ο επιλογέας δεδομένων για τα εγκλήματα έχει τη μορφή που παρουσιάζεται στην Εικόνα 5, στα δεξιά).

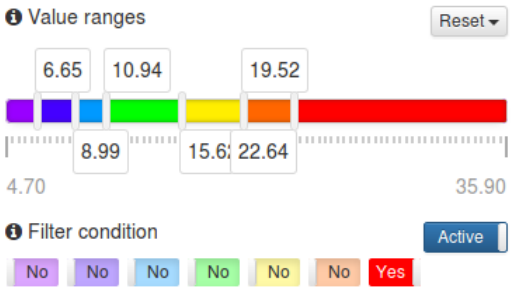


Εικόνα 6. Γραφήματα και χρωματισμός του χάρτη με βάση τα εγκλήματα που έγιναν μήνα Μάιο

Όπως φαίνεται από την Εικόνα 2 και την Εικόνα 6, με μόνη εξαίρεση τον επιλογέα δεδομένων, ο οποίος εμφανίζεται στην κορυφή του συρταριού, αν υπάρχει επιλογέας δεδομένων για το συγκεκριμένο μέγεθος (για τα εγκλήματα υπάρχει, λόγου χάρη, επιλογέας δεδομένων ενώ για την ανεργία δεν υπάρχει), η μορφή όλων των συρταριών που βρίσκονται στην ίδια μπάρα είναι ακριβώς η ίδια, ανεξάρτητα από το υπό εξέταση μέγεθος.

Όσον αφορά τα συρτάρια της δεξιάς μπάρας, εμφανίζεται στην κορυφή κάθε συρταριού ο επιλογέας δεδομένων για το συγκεκριμένο μέγεθος, προαιρετικά και εξειδικευμένα, όπως αναφέραμε, ανάλογα με το μέγεθος. Από κάτω εμφανίζεται μία κατανομή πίτας με το πλήθος και τα ποσοστά των tracts που εμπίπτουν σε 7 εύρη τιμών του μεγέθους, τα οποία ορίζονται παρακάτω στο συρτάρι. Κάτω από την κατανομή πίτας εμφανίζεται ένα ιστόγραμμα με την κατανομή των tracts σε μικρότερα εύρη τιμών, υποδιαιρέσεις των 7 ευρών τιμών. Κάτω από το ιστόγραμμα εμφανίζεται ένας χρωματικός ολισθητής εύρους πολλαπλών λαβών (multiple handle colored range slider), ο οποίος είναι ρυθμιζόμενος από το χρήστη, και ο οποίος χρησιμοποιείται, με τη μετακίνηση των λαβών, για τον ορισμό των 7 ευρών τιμών, κάθε εύρος από τα οποία αντιστοιχίζεται σε κάποιο χρώμα. Τα tracts που εμπίπτουν στο εύρος τιμών που χρωματίζεται με ένα συγκεκριμένο χρώμα θα χρωματιστούν με το χρώμα αυτό στο χάρτη. Κάτω από τον ολισθητή εύρους εμφανίζεται η συνθήκη φιλτραρίσματος για το υπό εξέταση μέγεθος. Η συνθήκη μπορεί να είναι ενεργή ή ανενεργή, το οποίο καθορίζει το αν το υπό εξέταση μέγεθος θα ληφθεί υπόψη στον υπολογισμό του καθολικού φίλτρου ή όχι, και συνίσταται περαιτέρω από τις υποσυνθήκες

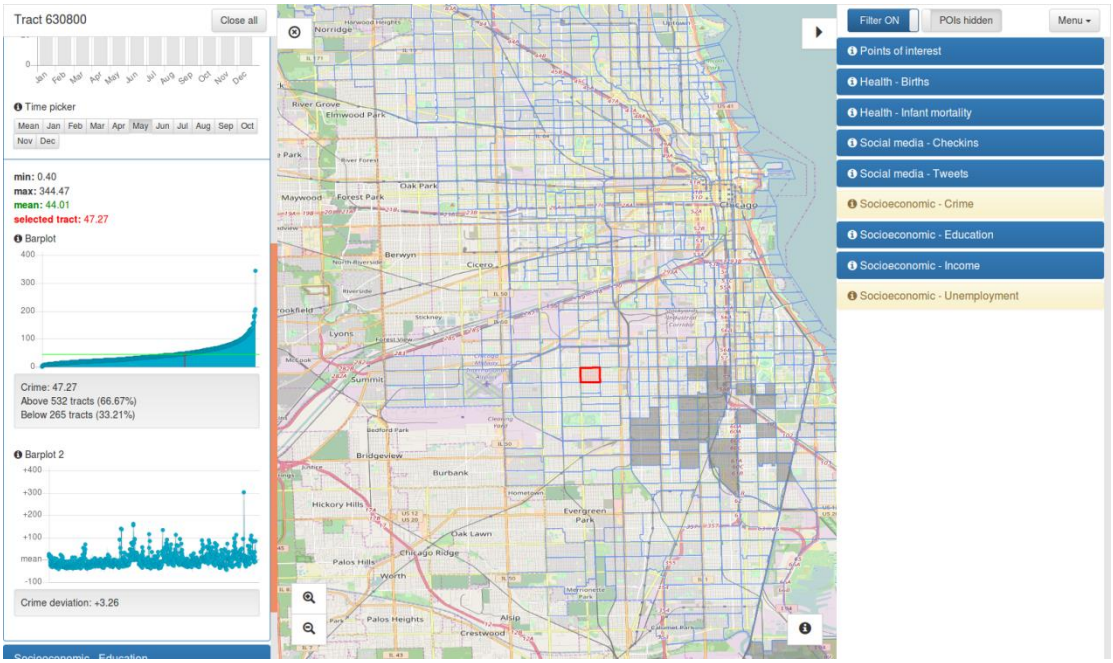
(με μία υποσυνθήκη για κάθε χρώμα) υπολογισμού, ή όχι, των επιμέρους χρωματισμένων ομαδοποιήσεων των tracts στο φίλτρο. Αν ένα χρώμα έχει την τιμή "No", τότε δεν θα συμπεριληφθεί στο φίλτρο. Αν έχει την τιμή "Yes", τότε θα συμπεριληφθεί στο φίλτρο. Στην Εικόνα 7 φαίνεται η διάταξη του χρωματικού ολισθητή εύρους και της συνθήκης φιλτραρίσματος για την περίπτωση της ανεργίας. Στο φιλτράρισμα θα συμπεριληφθεί μόνο το κόκκινο χρώμα, δηλαδή τα tracts τα οποία έχουν ανεργία 22.64% και πάνω.



Εικόνα 7. Χρωματικός ολισθητής εύρους και συνθήκη φιλτραρίσματος

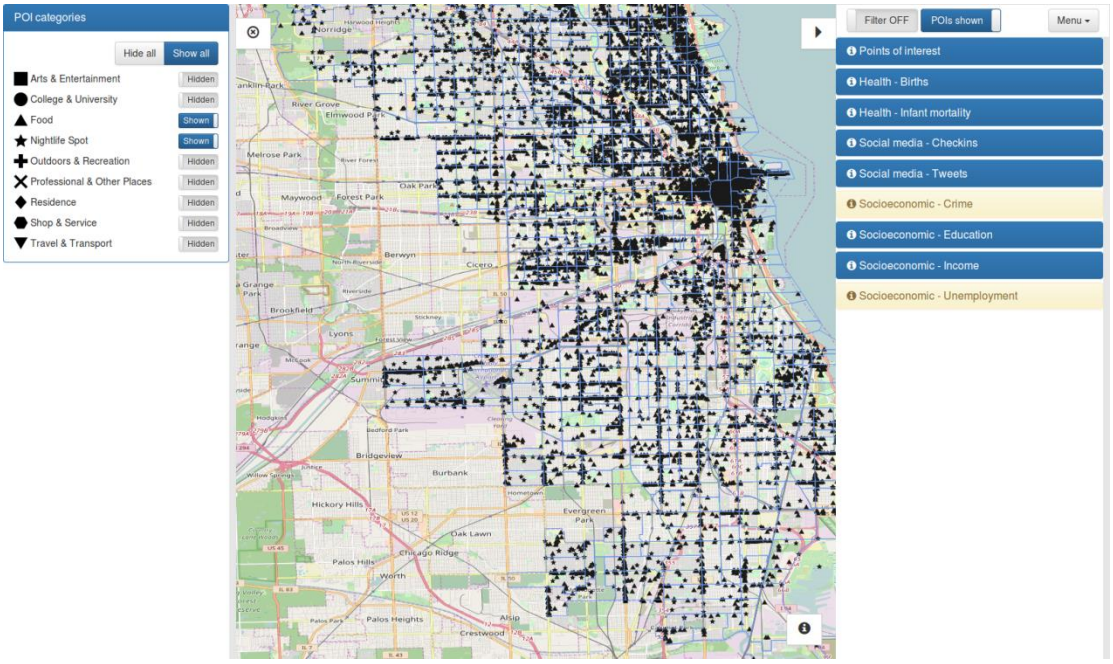
Όσον αφορά τα συρτάρια της αριστερής μπάρας, εμφανίζεται στην κορυφή κάθε συρταριού ο επιλογέας δεδομένων για το συγκεκριμένο μέγεθος, προαιρετικά και εξειδικευμένα, όπως αναφέραμε, ανάλογα με το μέγεθος. Από κάτω εμφανίζονται η ελάχιστη, η μέγιστη, και η μέση τιμή του μεγέθους για όλα τα tracts, καθώς και η τιμή του μεγέθους για το επιλεγμένο tract. Από κάτω εμφανίζεται το διάγραμμα των τιμών του μεγέθους για όλα τα tracts, σε αύξουσα κατά το μέγεθος σειρά. Με πράσινη γραμμή επισημαίνεται η μέση τιμή του μεγέθους και με κόκκινη γραμμή επισημαίνεται το επιλεγμένο tract. Σημειώνεται, επίσης, πόσα tracts βρίσκονται πάνω και κάτω από το επιλεγμένο tract όσον αφορά το συγκεκριμένο μέγεθος. Από κάτω εμφανίζεται το διάγραμμα της απόκλισης από τη μέση τιμή της τιμής του μεγέθους για όλα τα tracts, σε αύξουσα κατά τον κωδικό του tract σειρά. Το επιλεγμένο tract επισημαίνεται με κόκκινη γραμμή. Από κάτω σημειώνεται αριθμητικά η απόκλιση από τη μέση τιμή του μεγέθους για το επιλεγμένο tract.

Με το πάτημα του κουμπιού "Filter ON", που βρίσκεται στην κορυφή της δεξιάς μπάρας, ενεργοποιείται η λειτουργία φιλτραρίσματος. Λαμβάνεται η τομή των συνθηκών φιλτραρίσματος (η συνθήκη φιλτραρίσματος για κάποιο μέγεθος είναι η ένωση όλων των υποσυνθηκών φιλτραρίσματος, δηλαδή η άθροιση των tracts που ανήκουν στις χρωματικές ομαδοποιήσεις που έχουν "Yes") για τις ενεργοποιημένες συνθήκες φιλτραρίσματος, και τα tracts που ικανοποιούν όλες τις συνθήκες φιλτραρίσματος εμφανίζονται σκιασμένα με γκρι χρώμα στο χάρτη, όπως φαίνεται στην Εικόνα 8.



Εικόνα 8. Λειτουργία φιλτραρίσματος

Το CitySense παρέχει, τέλος, τη λειτουργία προβολής σημείων ενδιαφέροντος, τα οποία ανήκουν σε κατηγορίες που ενδιαφέρουν το χρήστη, πάνω στο χάρτη. Η συγκεκριμένη λειτουργία είναι ανεξάρτητη από τις λειτουργίες χρωματισμού χάρτη και φιλτραρίσματος, καθώς και από τη λειτουργία επιλογής tract. Μπορεί να χρησιμοποιηθεί, δηλαδή, σε συνδυασμό με οποιαδήποτε μορφή έχει η εφαρμογή την δεδομένη στιγμή. Στην Εικόνα 9 φαίνονται τα σημεία ενδιαφέροντος των κατηγοριών "Food" (επισημαίνονται με τρίγωνα) και "Nightlife Spot" (επισημαίνονται με αστέρια), οι οποίες έχουν ενεργοποιηθεί, με τα αντίστοιχα κουμπιά στη θέση "Shown", από τη λίστα των κατηγοριών σημείων ενδιαφέροντος στην αριστερή πλευρά, ενώ η λειτουργία προβολής σημείων ενδιαφέροντος ενεργοποιείται από τη δεξιά μπάρα με το πάτημα του κουμπιού "POIs shown".



Εικόνα 9. Προβολή σημείων ενδιαφέροντος επιλεγμένων κατηγοριών στο χάρτη

6 Τεχνικές Προκλήσεις και Καινοτόμα Χαρακτηριστικά

6.1 Οργάνωση ετερογενών συνόλων δεδομένων

Για την απόδοση της αίσθησης μιας πόλης, το CitySense πρέπει να ενσωματώσει και να οπτικοποιήσει μία ποικιλία συνόλων δεδομένων. Οι πηγές δεδομένων που ενσωματώνονται συνίστανται από δημογραφικά δεδομένα, δεδομένα κοινωνικών δικτύων, και δεδομένα σημείων ενδιαφέροντος. Η ποικιλόμορφη φύση αυτών των συνόλων δεδομένων απαιτεί διαφορετικό χειρισμό για την ενσωμάτωση των δεδομένων όσον αφορά στη χρονική πλευρά. Όπως φαίνεται στον Πίνακας 1:

- Τα ανοικτά δημογραφικά δεδομένα μπορούν να οπτικοποιηθούν τόσο στατικά (συνολικά στο χρόνο) όσο και χρονικά (ανά μήνα/μέρα/ώρα). Για παράδειγμα, τα δεδομένα υγείας και ανεργίας οπτικοποιούνται στατικά ενώ τα δεδομένα εγκληματικότητας οπτικοποιούνται χρονικά.
- Τα δεδομένα κοινωνικών δικτύων μπορούν να οπτικοποιηθούν στατικά, χρονικά, ή δυναμικά, παρόλο που παράγονται και συλλέγονται δυναμικά (σε πραγματικό χρόνο).
- Τα δεδομένα σημείων ενδιαφέροντος οπτικοποιούνται στατικά.

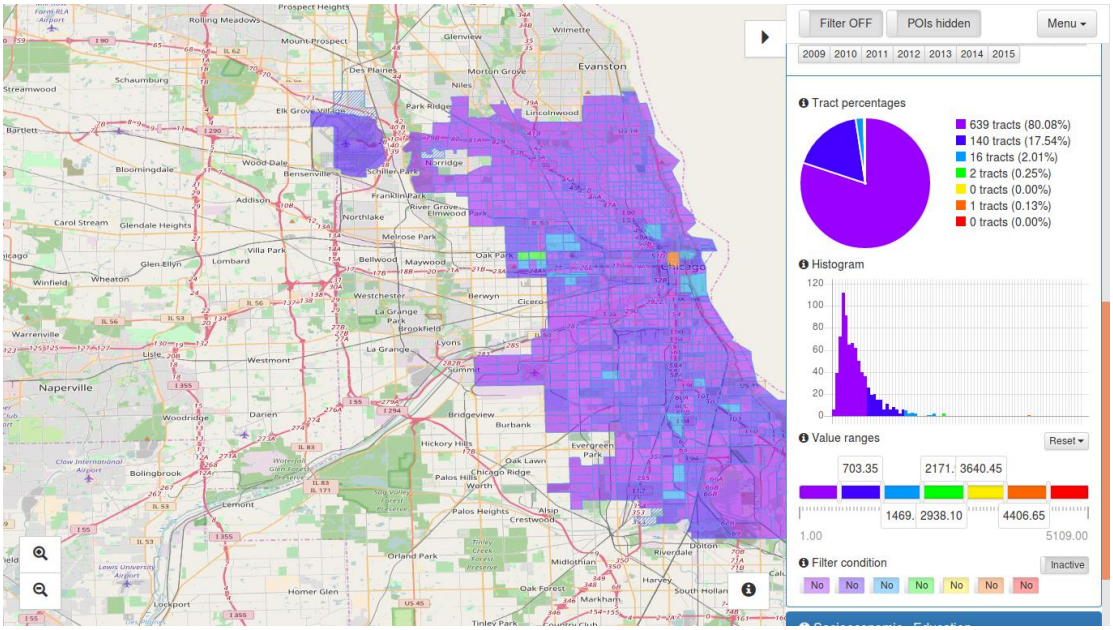
	Στατικά	Χρονικά
Ανοικτά δημογραφικά δεδομένα	✓	✓
Δεδομένα κοινωνικών δικτύων	✓	✓
Δεδομένα σημείων ενδιαφέροντος	✓	

Πίνακας 1. Ποικιλομορφία οπτικοποίησης συνόλων δεδομένων σε σχέση με το χρόνο

Η οργάνωση των δεδομένων που παρουσιάζεται στον Πίνακας 1 βοήθησε στην αντιμετώπιση της ποικιλομορφίας των δεδομένων, ώστε να επιτευχθεί συνεπής οπτικοποίηση και συνεπής χειρισμός των δεδομένων μέσα στην εφαρμογή.

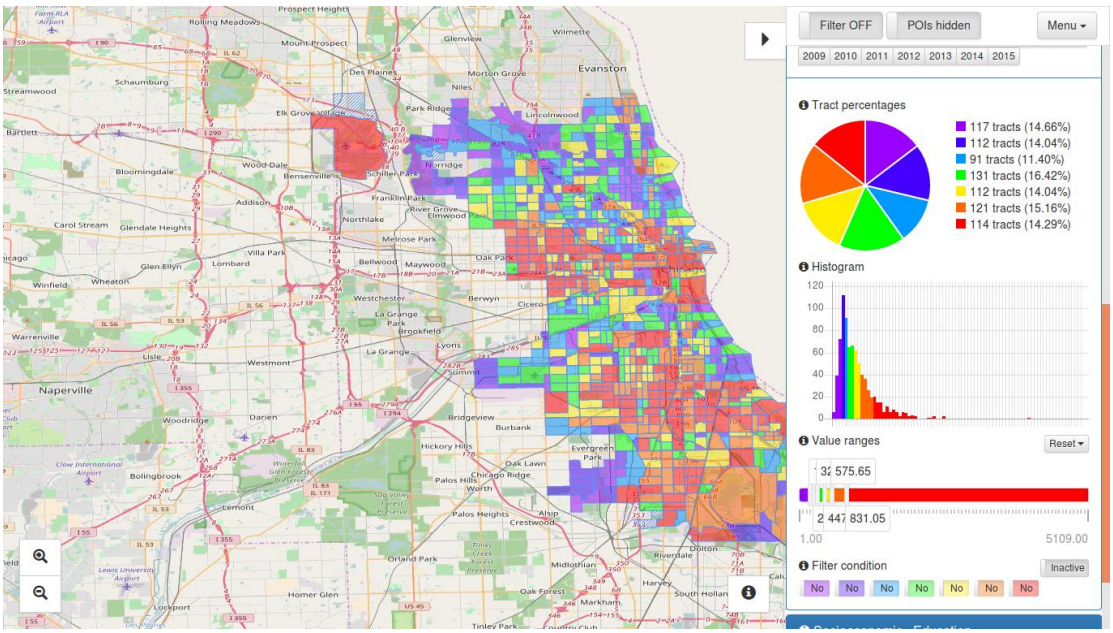
Ένα συναφές πρόβλημα είναι αυτό της αρχικοποίησης των ρυθμιζόμενων από το χρήστη χρωματικών ολισθητών εύρους (range sliders).

Ο στόχος μας ήταν η παροχή μίας εύλογης δυνατότητας χρωματισμού του χάρτη, ο οποίος θα βοηθάει τους χρήστες να συνάγουν συμπεράσματα για την πόλη. Παρείχαμε, έτσι, δύο επιλογές για την αρχικοποίηση. Η πρώτη επιλογή, αρχικοποίηση βασισμένη στην τιμή, χωρίζει τον ολισθητή με βάση ισαπέχουσες τιμές. Αυτή η προσέγγιση είναι, ωστόσο, ευαίσθητη σε δεδομένα με έντονα ακραίες τιμές ή με έντονη συγκέντρωση σε συγκεκριμένα εύρη. Η δεύτερη επιλογή παρέχει αρχικοποίηση βασισμένη σε ποστοστά, χωρίζει, δηλαδή, τον ολισθητή με βάση ίσες κατανομές ποσοστών. Αυτή η προσέγγιση, από την άλλη πλευρά, είναι ευαίσθητη στην ύπαρξη πολλών περιοχών με σχεδόν ίσες τιμές. Για παράδειγμα, η Εικόνα 10 παρουσιάζει την αρχικοποίηση με βάση την τιμή για δεδομένα εγκληματικότητας.



Εικόνα 10. Αρχικοποίηση με βάση την τιμή

Όπως μπορούμε να παρατηρήσουμε στο ιστόγραμμα που φαίνεται στην Εικόνα 10 (δεξιά), τα δεδομένα εγκληματικότητας καλύπτουν κατά κύριο λόγο ένα μικρό εύρος τιμών, ανάμεσα στο 1 και το 1469, με αποτελέσμα ένα χάρτη χρωματισμένο κυρίως με δύο χρώματα (βιολετί και βαθύ μπλε) στην Εικόνα 10 (αριστερά). Για την αντιμέτωπιση αυτού του προβλήματος χρησιμοποιούμε την αρχικοποίηση με βάση τα ποσοστά, η οποία παρουσιάζεται στην Εικόνα 11.



Εικόνα 11. Αρχικοποίηση με βάση τα ποσοστά

Ο χρωματισμός του χάρτη που προκύπτει είναι εμφανώς βελτιωμένος στην Εικόνα 11 (αριστερά). Όπως μπορούμε να παρατηρήσουμε, ωστόσο, στα ποσοστά των περιοχών που παρουσιάζονται στην Εικόνα 11 (δεξιά), οι κατανομές δεδομένων εγκληματικότητας δεν είναι χωρισμένες ισόποσα, καθώς υπάρχουν περιοχές που έχουν τις ίδιες σχεδόν τιμές,

αναφορικά με το βήμα εύρους (range step). Για αυτό το λόγο δεν μπορούν οι περιοχές να κατενεμηθούν εντελώς ομοιόμορφα.

6.2 Συλλογή δεδομένων μίας περιοχής

Ο CityProfiler συλλέγει δεδομένα σημείων ενδιαφέροντος από μία αστική περιοχή εκτελώντας κλήσεις σε προγραμματιστικές διεπαφές υπηρεσιών, όπως είναι το Google Places και το Foursquare, οι οποίες θέτουν περιορισμούς. Μία αφελής τακτική συλλογής των σημείων ενδιαφέροντος, στην κλίμακα μίας ολόκληρης πόλης, δεν θα ήταν εφικτό να συλλέξει όλο το πλήθος των σημείων ενδιαφέροντος, αλλά μόνο ένα μικρό ποσοστό του, με βάση τους κανόνες που τίθενται από την πηγή. Το CitySense αντιμετωπίζει αυτό το θέμα με τον τεμαχισμό της περιοχής σε μικρότερα κομμάτια προκαταβολικά. Συγκεκριμένα, η πόλη χωρίζεται σε τετράγωνα με μήκος και πλάτος 0.03 μοιρών, πριν ξεκινήσει η συλλογή των σημείων ενδιαφέροντος. Στην περίπτωση που η μέθοδος αυτή δεν συλλέξει όλα τα σημεία ενδιαφέροντος, τότε χρησιμοποιείται αναδρομή.

Το CitySense συλλέγει, επίσης, δεδομένα κοινωνικών δικτύων για την πόλη σε πραγματικό χρόνο. Για το σκοπό αυτό, το CityProfiler εκτελεί συλλογή των tweets σε πραγματικό χρόνο με χρήση της προγραμματιστικής διεπαφής Twitter Streaming, θέτοντας ως παράμετρο φιλτραρίσματος ένα κουτί με τα γεωγραφικά όρια της τοποθεσίας, το οποίο περικλείει την πόλη. Συλλέγονται μόνο τα γεωγραφικά εντοπισμένα tweets (τα οποία δημοσιεύονται μαζί με το γεωγραφικό μήκος και πλάτος τους). Για τη συλλογή των κοινωνικών check-in και των σημείων ενδιαφέροντος από τα οποία δημοσιεύονται, το CityProfiler εκτελεί μία κλήση στην προγραμματιστική διεπαφή του Foursquare κάθε φορά που ένα tweet περιέχει ένα σύνδεσμο Swarm (εφαρμογή που επιτρέπει στους χρήστες να μοιράζονται την τοποθεσία τους εντός του κοινωνικού δικτύου τους). Με αυτόν τον τρόπο, η εφαρμογή συλλέγει χρονική πληροφορία αναφορικά με τα γεωγραφικά εντοπισμένα tweets, συμπεριλαμβανομένων των hashtags και των check-in που δημοσιεύονται στα σημεία ενδιαφέροντος της πόλης.

6.3 Θέματα υλοποίησης και απόδοσης

Έπρεπε να ληφθούν αρκετές αποφάσεις σχετικά με την υλοποίηση, έτσι ώστε η εφαρμογή να τρέχει αποδοτικά. Έπρεπε η εφαρμογή να είναι ελαφριά σε σχέση με την κατανάλωση μνήμης και επεξεργαστικής ισχύος, αλλά και αποκρίσιμη σε σχέση με την εμπειρία χρήσης του τελικού χρήστη.

Σε ορισμένα τμήματα της εφαρμογής χρειάζεται να εμφανιστεί στην οθόνη την ίδια στιγμή ένας μεγάλος αριθμός γεωμετριών, συγκεκριμένα δεκάδες χιλιάδες σημεία ενδιαφέροντος. Η επιλογή του χειρισμού κάθε γεωμετρίας ως ξεχωριστής οντότητας και η ξεχωριστή σχεδίασή της στο χάρτη θα απαιτούσε αρκετή μνήμη και επεξεργαστική ισχύ, ειδικά κατά το ζουμ μέσα και έξω στο χάρτη. Η προσέγγιση που ακολουθήθηκε βασίζεται στη σχεδίαση των σχετικών γεωμετριών ως μίας στρώσης εικόνας, η οποία περιέχει όλες τις γεωμετρίες. Έγινε χρήση του GeoServer (βλέπε Εικόνα 1) για την παραγωγή και την αποστολή στρώσεων εικόνας. Για επιπλέον απόδοση χρησιμοποιήθηκε η ενσωματωμένη στο GeoServer λειτουργία κρυφής μνήμης (caching) για τις στρώσεις εικόνας. Με αυτόν τον τρόπο μπορούν να χρησιμοποιηθούν σε επόμενα αιτήματα στρώσεις εικόνας που έχουν ήδη δημιουργηθεί προηγουμένως.

Οι απαιτήσεις της εφαρμογής συμπεριλαμβάνουν αθροιστικά ερωτήματα σε δεδομένα, τόσο κατά τη χωρική όσο και κατά τη χρονική διάσταση. Τέτοια ερωτήματα απαιτούν χρόνο, αν εκτελεστούν σε ακατέργαστα δεδομένα, με αποτέλσμα τη μείωση της αποκρισιμότητας για τον τελικό χρήστη. Για την αποφυγή των χρονοβόρων πράξεων κατά το χρόνο εκτέλεσης, χρησιμοποιείται ένα στάδιο προεπεξεργασίας. Ο σχεδιασμός της βάσης δεδομένων για τα

προεπεξεργασμένα δεδομένα καθορίστηκε από τις κρίσιμες περιπτώσεις χρήσης για τον τελικό χρήστη μέσω της διεπαφής χρήστη. Για παράδειγμα, ο χρήστης μπορεί να θέσει ερωτήματα για δεδομένα check-in, αθροισμένα ανά περιοχή, σχετιζόμενα με μία συγκεκριμένη κατηγορία σημείων ενδιαφέροντος και μία συγκεκριμένη μέρα της εβδομάδας. Τα ακατέργαστα δεδομένα check-in περιέχουν τις γεωγραφικές συντεταγμένες του σημείου ενδιαφέροντος, την κατηγορία του σημείου ενδιαφέροντος, όπως και την ημερομηνία και ώρα του check-in, σε δύο πίνακες. Οι γεωμετρίες των περιοχών αποθηκεύονται, επίσης, σε ξεχωριστό πίνακα. Ένα τέτοιο ερώτημα δεν μπορεί να εκτελεστεί ακαριαία. Κατά το στάδιο προεπεξεργασίας, οι συντεταγμένες των σημείων ενδιαφέροντος αντιστοιχίζονται στις περιοχές από τις οποίες περικλείονται, οι μέρες της εβδομάδας εξάγονται από την ημερομηνία, και εκτελείται η άθροιση ανά περιοχή και μέρα της εβδομάδας. Τα αποτελέσματα της προεπεξεργασίας αποθηκεύονται σε πίνακες της βάσης δεδομένων. Με αυτόν τον τρόπο επιτυγχάνονται αποδοτικά ερωτήματα για check-in σε σημεία ενδιαφέροντος συγκεκριμένης κατηγορίας σε συγκεκριμένη μέρα της εβδομάδας. Χρησιμοποιούνται ξεχωριστοί πίνακες για τις διαφορετικές περιπτώσεις χρονικής ανάλυσης στη διάσταση του χρόνου. Υπάρχουν, δηλαδή, ξεχωριστοί πίνακες για τα χρόνια, τους μήνες, τις μέρες της εβδομάδας, τις ώρες της ημέρας. Ένα άλλο μέτρο βελτιστοποίησης προς την ίδια κατεύθυνση είναι η ανάθεση κοστοβόρων υπολογισμών στο στάδιο αρχικοποίησης των λειτουργικών τμημάτων (services) της εφαρμογής. Το συγκεκριμένο μέτρο επηρεάζει το χρόνο εκκίνησης της εφαρμογής, αλλά επιταχύνει τα ερωτήματα στο χρόνο εκτέλεσης.

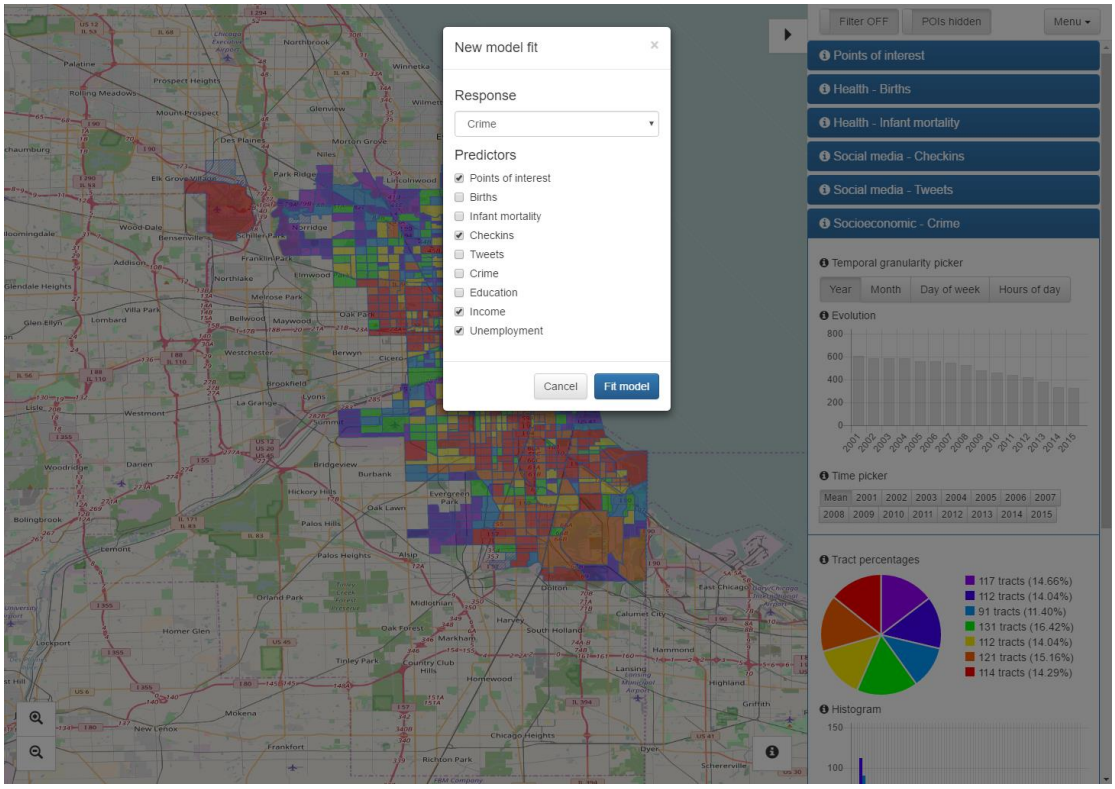
6.4 Γραμμικό μοντέλο πρόβλεψης

Είναι συχνές οι περιπτώσεις που τα σύνολα δεδομένων δεν είναι ανεξάρτητα μεταξύ τους. Για παράδειγμα, η παιδική θνησιμότητα είναι πολύ πιθανό ότι συνδέεται με το εισόδημα, και είναι αυξημένη σε περιοχές με χαμηλό εισόδημα. Ένας τρόπος να προβλέπει κανείς τιμές μιας μεταβλητής (response) με βάση τις αντίστοιχες τιμές άλλων μεταβλητών (predictors) βασίζεται στην μέθοδο των ελάχιστων τετραγώνων και στην εύρεση του κατάλληλου γραμμικού μοντέλου (linear model). Παρόλο που είναι συχνά απλουστευτικά, τα γραμμικά μοντέλα έχουν το πλεονέκτημα ότι είναι εύκολα ερμηνεύσιμα, όπως θα εξηγηθεί στην συνέχεια.

Το CitySense υποστηρίζει την δημιουργία γραμμικών μοντέλων για οποιοδήποτε από τα σύνολα δεδομένων που χειρίζεται. Υπάρχουν δύο λόγοι για την δημιουργία τέτοιων μοντέλων:

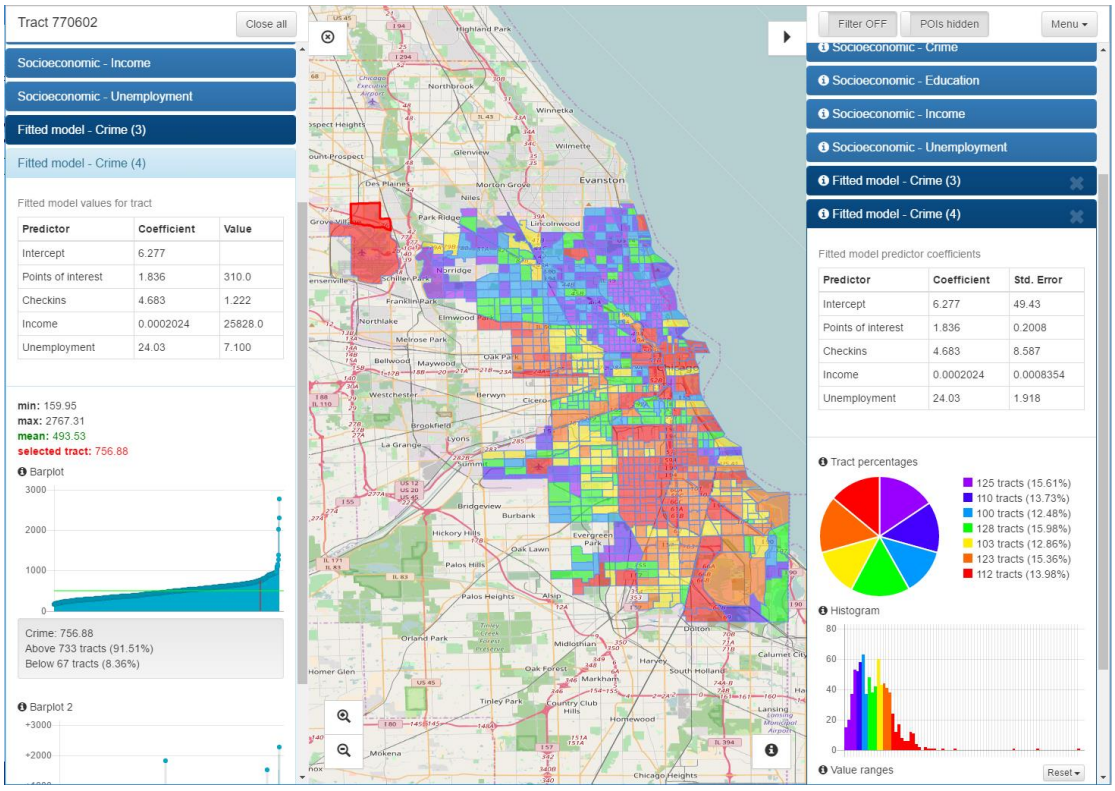
- Επιτρέπουν μια «διερευνητική ανάλυση» των δεδομένων. Μέσα από την σύγκριση των προβλεπόμενων τιμών με τις πραγματικές γίνεται δυνατόν να διερευνηθούν πιθανοί συσχετισμοί μεταξύ των μεταβλητών, πχ. ότι τα εγκλήματα συνδέονται με το εισόδημα και την ανεργία.
- Επιτρέπουν την εκτίμηση μιας τιμής που δεν υπάρχει για κάποια περιοχή, επειδή ακριβώς η τιμή αυτή μπορεί να προβλεφθεί με βάση ττις τιμές των predictors για την συγκεκριμένη περιοχή.

Σαν παράδειγμα θεωρούμε τα εγκλήματα (crimes). Από το μενού της εφαρμογής μπορούμε να δημιουργούμε κατά βούληση γραμμικά μοντέλα (επιλογή "New model fit") με το crimes ως response και για οποιοδήποτε συνδυασμό predictors, όπως φαινεται στην Εικόνα 12. Συγκεκριμένα στην Εικόνα 12 έχουμε επιλέξει σαν predictors του crime από την μια τα income και unemployment και από την άλλη τα checkins και points of interest με την λογική ότι στα πιο πολυσύχναστα σημεία θα έχουμε και περισσότερα εγκλήματα.



Εικόνα 12: Δημιουργία γραμμικού μοντέλου

Το αποτέλεσμα του γραμμικού μοντέλου (οι προβλέψεις για τις τιμές του crime) φαινονται στην Εικόνα 13, η οποία σε σύγκριση με τα πραγματικά δεδομένα για το crime επιβεβαιώνει σε γενικές γραμμές την σχέση του crime με τους συγκεκριμένους predictors.



Εικόνα 13: Απεικόνιση του γραμμικού μοντέλου

Παρατηρούμε ότι στην Εικόνα 12 δεν υπάρχει τιμή για το πάνω αριστερό tract, που για αυτό τον λόγο εμφανίζεται γραμμοσκιασμένο. Το ίδιο tract έχει τιμή και εμφανίζεται κόκκινο στην Εικόνα 13. Επειδή είναι και επιλεγμένο (κόκκινο περίγραμμα), τα στοιχεία για το συγκεκριμένο tract όπως προκύπτουν από το γραμμικό μοντέλο εμφανίζονται στο αριστερό συρτάρι. Σύμφωνα με την θεωρία, η τιμή του crimes για το εν λόγω tract με αριθμο 770602 είναι 756.88 και προκύπτει ως γραμμικός συνδυασμός των τιμών των predictors για το συγκεκριμένο tract:

6.277 + (1.836 * 310) + (4.683 *1.222) + (0.0002024 * 25828) + (24.03 * 7.1) = 757

Οι συντελεστές (coefficients) των τιμών προκύπτουν από την μέθοδο των ελάχιστων τετραγώνων και καθορίζουν το συγκεκριμένο γραμμικό μοντέλο.

6.5 Ανάλυση Άποψης

Ως Ανάλυση Συναισθήματος (Sentiment Analysis) ορίζεται η υπολογιστική μελέτη απόψεων, συναισθημάτων, στάσεων και, γενικότερα, υποκειμενικών καταστάσεων που εκφράζονται σε κείμενα [47]. Δοθέντος ενός κειμένου ή τμήματός του (π.χ. μια κριτική για ένα εστιατόριο), βασικός στόχος είναι η κατηγοριοποίησή του σε θετικό ή αρνητικό με βάση το σημασιολογικό προσανατολισμό των συναισθημάτων, απόψεων κτλ., που εκφράζονται σε αυτό [48, 49]. Ιστότοποι, όπως το Foursquare ή το Amazon, οι οποίοι φιλοξενούν κριτικές πελατών για σημεία ενδιαφέροντος ή προϊόντα αντίστοιχα, παρέχουν συνήθως αυτήν την πληροφορία με τη μορφή βαθμολογίας συνολικά. Αυτή η βαθμολογία βασίζεται συνήθως στις αξιολογήσεις των πελατών με τη μορφή βαθμολόγησης σε συγκεκριμένη γενική συνολική κλίμακα (ratings) και όχι στις κειμενικές κριτικές.

Παρόλο που η συνολική βαθμολόγηση ενός σημείου ενδιαφέροντος έχει μια πληθώρα χρήσιμων εφαρμογών, δεν παρέχει πληροφορία σχετικά με τα επιμέρους χαρακτηριστικά του για τα οποία εκφράζεται άποψη στις κριτικές των πελατών. Οι κειμενικές κριτικές είναι σημαντικές, καθώς μπορούν να παρέχουν πληροφορία που βοηθά να κατανοήσει κανείς τους λόγους πίσω από τη βαθμολογία [50, 51] π.χ. για ποιό λόγο ένα εστιατόριο βαθμολογείται με τρία και όχι πέντε αστέρια; Τι είναι αυτό που δεν ικανοποιεί τους πελάτες; Για παράδειγμα στην περίπτωση του CitySense, θα ήταν χρήσιμο για έναν κάτοικο ή επισκέπτη μιας πόλης να μπορεί να φιλτράρει συγκεκριμένα σημεία ενδιαφέροντος (π.χ. εστιατόρια, κέντρα διασκέδασης) όχι μόνο με βάση τη συνολική αξιολόγησή τους, αλλά εστιάζοντας σε συγκεκριμένα χαρακτηριστικά όπως οι τιμές (value for money).

Συνεπώς, υπάρχει η ανάγκη για πιο λεπτομερείς (fine-grained) προσεγγίσεις, και συγκεκριμένα, για ανάλυση άποψης για οντότητες και χαρακτηριστικά αυτών (Aspect-based Sentiment Analysis- ABSA) εστιάζοντας, δηλαδή, σε συγκεκριμένα χαρακτηριστικά οντοτήτων ενδιαφέροντος και όχι μόνο στη γενική κατηγοριοποίηση ενός κειμένου σε θετικό ή αρνητικό. Για παράδειγμα, δοθέντος ενός κειμένου ή τμήματός του (π.χ. μια κριτική για ένα εστιατόριο), στόχος είναι ο εντοπισμός των επιμέρους χαρακτηριστικών του εστιατορίου (π.χ. φαγητό, εξυπηρέτηση, τιμή, ατμόσφαιρα κτλ.) και των απόψεων που εκφράζονται για τα χαρακτηριστικά αυτά [52]. Μια τέτοια μέθοδος μπορεί να αναλύσει μεγάλες ποσότητες μη δομημένων κειμένων και να εξάγει πληροφορία που δεν περιλαμβάνεται στις γενικές βαθμολογίες/αξιολογήσεις (ratings) των χρηστών. Ανάλογα με την προσέγγιση, ο όρος χαρακτηριστικό (aspect) υποδηλώνει διαφορετικούς τύπους πληροφορίας και υπάρχουν διάφοροι τρόποι αναπαράστασης τους συγκεκριμένου προβλήματος:

- Προκαθορισμένες γενικές κατηγορίες (έννοιες) π.χ. FOOD, SERVICE [e.g. 51, 53].

- Πτυχές ή χαρακτηριστικά [54] που δηλώνουν επιμέρους μέρη/συστατικά (π.χ. food, pizza, pasta) ή ιδιότητες (π.χ. price, taste) της οντότητας ενδιαφέροντος και των επιμέρους συστατικών της [55].
- Συνδυασμός οντότητας (Entity) και ιδιότητας (Attribute) της (E#A pair), όπου το E μπορεί να είναι η ίδια η οντότητα ενδιαφέροντος (π.χ. εστιατόριο) ή κάποιο επιμέρους μέρος/συστατικό της (π.χ. φαγητό) ή κάποια άλλη σχετική οντότητα (π.χ. ανταγωνιστική επιχείρηση), και το A αντιστοιχεί σε ιδιότητες του E [56, 57]. Τα E και A είναι οντολογικές έννοιες-κατηγορίες που προκαθορίζονται με βάση το εκάστοτε θεματικό πεδίο (domain).

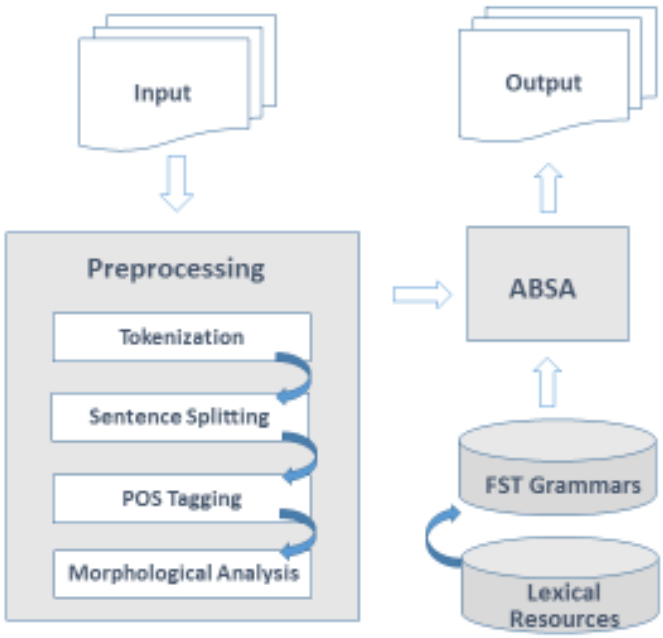
Αρκετές μέθοδοι ABSA έχουν προταθεί για διάφορους τομείς, όπως ηλεκτρονικά είδη [54], ταινίες [58], εστιατόρια [53], και μπυραρίες/παμπ [51], μεταξύ άλλων. Οι διαθέσιμες μέθοδοι μπορούν να ταξινομηθούν σε εκείνες που υιοθετούν λύσεις ανεξαρτήτως θεματικού πεδίου (domain indepedent) [59], καθώς και σε εκείνες που χρησιμοποιούν domain-specific γνώσεις για να βελτιώσουν τα αποτελέσματά τους [58]. Ορισμένες προσεγγίσεις αντιμετωπίζουν την κατηγοριοποίηση άποψης (sentiment classification) και την εξαγωγή χαρακτηριστικών (aspect detection) ως δύο ξεχωριστά προβλήματα [60, 61], ενώ άλλες τα μοντελοποιούν ως ενιαίο πρόβλημα [62, 63]. Μεταξύ των πιο κοινών τεχνικών που χρησιμοποιούνται είναι οι προσεγγίσεις που βασίζονται σε σχέσεις συχνότητας [64] ή συντακτικές σχέσεις [65], σε τεχνικές μηχανικής μάθησης [63, 66], μοντέλα Θεμάτων (Topic Modelling) [50, 61, 62] και σε νευρωνικά δίκτυα [67]. Οι βασικοί τύποι χαρακτηριστικών που χρησιμοποιούνται συνοψίζονται παρακάτω:

- Λεκτικά χαρακτηριστικά π.χ. n-grams, Token shape (π.χ. κεφαλαίοι χαρακτήρες)
- Μορφο-συντακτικά χαρακτηριστικά π.χ. μέρος του λόγου (POS), Δέντρα Εξάρτησης (Dependency trees)
- Σημασιολογικά χαρακτηριστικά π.χ. συστάδες λέξεων, σημασιολογικές εξαρτήσεις
- Λεξιλογικά χαρακτηριστικά π.χ. sentiment polarity lexica , WordNet
- Διανυσματικές αναπαραστάσεις λέξεων (Word Vector Representations) π.χ. Word2Vec [68], GloVe [69]

Για τις ανάγκες του CitySense υιοθετούμε μια γενική λύση (domain indepedent) που δε λαμβάνει υπόψιν συγκεκριμένο θεματικό πεδίο (π.χ. εστιατόρια), έτσι ώστε να μπορεί να έχει εφαρμογή σε πέραν του ενός τύπου σημείων ενδιαφέροντος (π.χ. food, nightlife spot, arts and entertainment, κ.ά.). Συγκεκριμένα, εστιάζουμε στο πως αξιολογούν επισκέπτες/πελάτες δύο βασικά χαρακτηριστικά ενός σημείου ενδιαφέροντος:

- PRICE (value for money)
- SERVICE (εξυπηρέτηση)

Το σύστημα που αναπτύξαμε είναι μια νομοθετική προσέγγιση που βασίζεται σε λεξικο-συντακτικές σχέσεις και ενσωματώνει γραμματικές προτύπων (pattern grammars) και λεξιλογικούς πόρους που σχεδιάστηκαν για τις ανάγκες της εφαρμογής. Η αρχιτεκτονική του συστήματος απεικονίζεται στο παρακάτω σχήμα:



Εικόνα 14: Αρχιτεκτονική συστήματος ανάλυσης άποψης

Σε πρώτη φάση γίνεται προεπεξεργασία των κειμένων εισόδου. Η διαδικασία της προεπεξεργασίας ακολουθεί την τυπική αλυσίδα επεξεργασίας κειμένου, η οποία περιλαμβάνει ένα σύνολο εργαλείων[1] που επιτελούν τις ακόλουθες εργασίες:

- Αναγνώριση λεκτικών μονάδων (Tokenization)
- Αναγνώριση ορίων-διαχωρισμός προτάσεων (Sentence Splitting)
- Μορφοσυντακτικός χαρακτηρισμός λεκτικών μονάδων (Part-Of-Speech tagging)
- Μορφολογική ανάλυση λεκτικών μονάδων (Morphological analysis)

Το αποτέλεσμα της προεπεξεργασίας αποτελεί είσοδο για τον αναλυτή γνώμης, ο οποίος σε ένα πρώτο στάδιο εντοπίζει στα κείμενα εισόδου υποψήφια χαρακτηριστικά (aspects) για τα οποία εκφράζονται απόψεις με βάση τα σχετικά λεξικά. Στη συνέχεια, μια σειρά αρθρωμάτων (modules), τα οποία περιλαμβάνουν γλωσσολογικούς κανόνες που μοντελοποιούν λεξικοσυνταντκές σχέσεις επιφανειακής δομής αποσαφηνίζουν το περιβάλλον γύρω από τα υποψήφια χαρακτηριστικά-στόχους και απόψεις. Σε ένα τρίτο στάδιο γίνεται η σύνδεση των χαρακτηριστικών με τις απόψεις που εκφράζονται για αυτά.

Ο αναλυτής γνώμης είναι μια παραλλαγή του συστήματος που παρουσιάζεται στο [70]. Το σύστημα υλοποιείται μέσω μιας ακολουθίας από πεπερασμένους μεταγραφείς (finite state transducers) με είσοδο την αναπαράσταση του κειμενικού υλικού (ακολουθία από εμπλουτισμένες γλωσσολογικά δομές των λεκτικών μονάδων κάθε πρότασης). Η έξοδος του συστήματος είναι επισημειώσεις τριπλέτες που περιλαμβάνουν τους εξής τύπους πληροφορίας (Εικόνα 15):

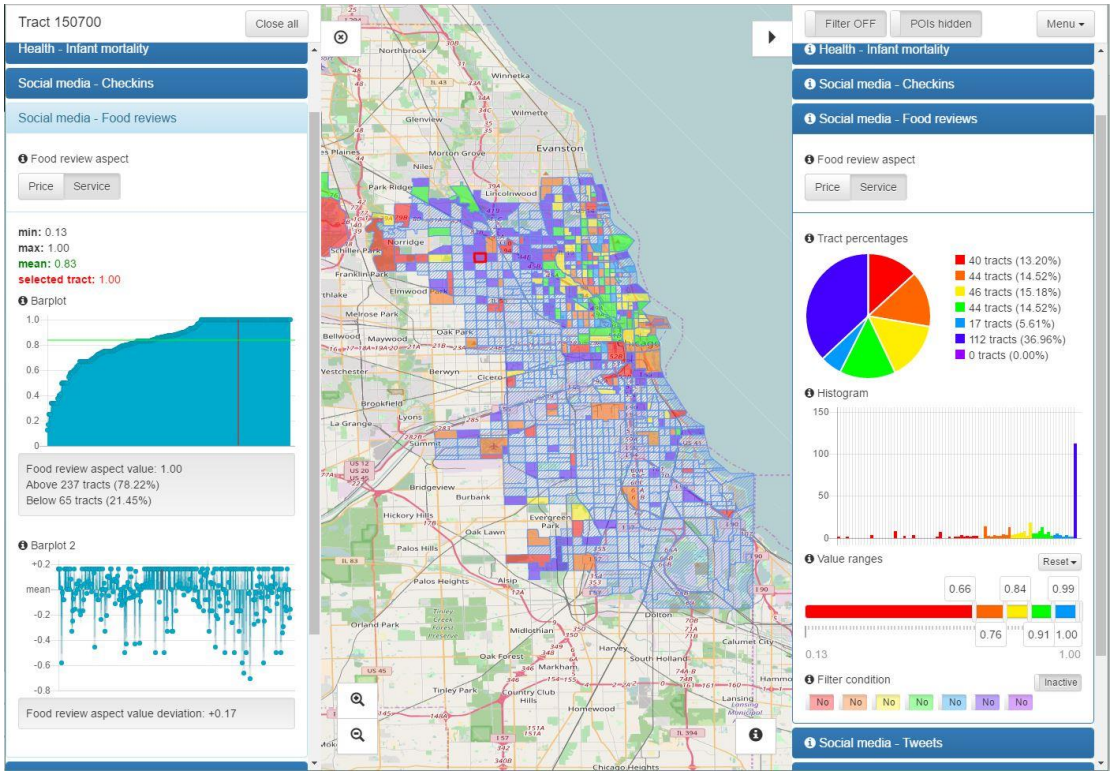
- χαρακτηριστικό (aspect) π.χ. PRICE
- προσανατολισμός άποψης για το χαρακτηριστικό αυτό π.χ. positive
- το σχετικό απόσπασμα κειμένου π.χ. "fair prices"

[1] Για την προεπεξεργασία χρησιμοποιούνται τα εργαλεία του συστήματος εξαγωγής πληροφορίας ANNIE μέσω του περιβάλλοντος του GATE (http://gate.ac.uk/sale/tao/splitch6.html#chap:annie)

Significant service and fair prices. I would absolutely suggest Morning Noon Night Plumbing & Sewer!	positive#PRICE#51#fair prices positive#SERVICE#50#Significant service

Εικόνα 15: Παράδειγμα εξόδου του συστήματος ABSA

Όπως φαίνεται και στην παραπάνω εικόνα, η εξαγωγή άποψης γίνεται σε επίπεδο έκφρασης για κάθε κείμενο εισόδου. Προκειμένου να ενσωματωθούν τα αποτελέσματα της ανάλυσης στην εφαρμογή, γίνεται συνάθροιση (aggregation) των αποτελεσμάτων του συστήματος ABSA, αρχικά, για καθένα από τα Points of Interest και στη συνέχεια για κάθε tract. Το αποτέλεσμα της συνάθροισης είναι ο μέσος όρος του sentiment (σε κλίμακα 0 έως 1) για κάθε poi και κάθε tract δεδομένου ενός χαρακτηριστικού (PRICE ή SERVICE) και ο χάρτης χρωματίζεται αναλόγως, όπως φαίνεται στην παρακάτω εικόνα:



Εικόνα 16: Χρωματισμός χάρτη με βάση σχόλια για food service

6.6 Προστασία της ιδιωτικότητας

Σήμερα, πολλοί οργανισμοί, επιχειρήσεις και δημόσιες υπηρεσίες συλλέγουν και διαχειρίζονται ένα τεράστιο αριθμό από προσωπικές πληροφορίες. Ένα τυπικό παράδειγμα τέτοιων πληροφοριών είναι τα κλινικά δεδομένα των νοσοκομείων, τα ερωτήματα σε μηχανές αναζητήσεις, τα δεδομένα πιστωτικών καρτών, κ.α. Οι ιδιοκτήτες των δεδομένων συχνά θέλουν να μοιραστούν ή να δημοσιεύσουν τα δεδομένα αυτά, με σκοπό να εξάγουν κάποια πληροφορία από αυτά, ή για να συνεισφέρουν σε κάποια επιστημονική έρευνα. Τέτοιες δημοσιεύσεις μπορεί να περιέχουν σημαντικά και ουσιώδη αποτελέσματα, αλλά την ίδια στιγμή απειλούν την ιδιωτικότητα των ατόμων που συσχετίζονται με αυτά τα δεδομένα. Για να αντιμετωπίσουμε αυτό το πρόβλημα, έχουν προταθεί στην επιστημονική βιβλιογραφία, διάφορες μέθοδοι ανωνυμοποίησης δεδομένων. Η ανωνυμοποίηση

δεδομένων μετατρέπει τα αρχικά δεδομένα σε μία νέα μορφή , όπου οι χρήστες δεν μπορούν να αναγνωριστούν, διατηρώντας όμως όσον το δυνατόν περισσότερες πληροφορίες από τα αρχικά δεδομένα.

Καθώς η τεχνολογία εξελίσσεται και αρχίζει να επηρεάζει όλο και περισσότερες πτυχές της ζωής μας, ένας μεγάλος αριθμός δεδομένων που περιλαμβάνει σημαντικές προσωπικές πληροφορίες δημιουργείται και αποθηκεύεται καθημερινά. Ο διαμοιρασμός και η εκμετάλλευση αυτών των δεδομένων αποτελεί μία πολύ σημαντική πηγή νέων πληροφοριών, όμως από την άλλη πλευρά απειλεί την ιδιωτικότητα των χρηστών.

Πολλές τεχνικές ανωνυμοποίησης δεδομένων έχουν υλοποιηθεί τα τελευταία χρόνια, οι οποίες μετατρέπουν τα αρχικά δεδομένα σε μία νέα μορφή, όπου οι πληροφορίες που μπορούν να ταυτοποιήσουν μια πληροφορία για κάποιο χρήστη έχουν αφαιρεθεί. Ταυτόχρονα όμως, τα δεδομένα παραμένουν χρήσιμα για ανάλυση και εξαγωγή σημαντικών πληροφοριών. Οι μέθοδοι ανωνυμοποίησης δεδομένων προχωράνε πιο πέρα, από την απλή αφαίρεση άμεσων αναγνωριστικών στοιχείων, όπως είναι το όνομα, ή το ΑΦΜ, αποκρύπτουν και δευτερεύουσες πληροφορίες, όπως την ημέρα γέννησης, τον ταχυδρομικό κώδικα, οι οποίες μπορούν να διασταυρωθούν με άλλες εξωτερικές πηγές, όπως οι εκλογικοί κατάλογοι και να οδηγήσουν στον επαναπροσδιορισμό ενός χρήστη. Αυτές οι δευτερεύουσες πληροφορίες, ονομάζονται ψευδοαναγνωριστικά. Διαφορετικές τεχνικές ανωνυμοποίησεις δεδομένων παρέχουν διαφορετικές στατιστικές εγγυήσεις για τα ανωνυμοποιημένα δεδομένα για να αντιμετωπίσουν διαφορετικά σενάρια επιθέσεων. Για παράδειγμα η μέθοδος k-anonymity εγγυάται ότι κάθε ανωνυμοποιημένη εγγραφή, δεν θα μπορεί να αναγνωριστεί για από κ-1 εγγραφές λαμβάνοντας υπόψιν τα ψευδοαναγνωριστικά. Η μέθοδος l-diversity δεν θα επιτρέψει σε έναν επιτιθέμενο να συνδέσει μία εγγραφή με λιγότερες από l well-represented τιμές η μέθοδος differential privacy μειώνει την επιρροή από την απουσία ή την παρουσία μίας εγγραφής σε ένα αποτέλεσμα ερωτήματος στα ανωνυμοποιημένα δεδομένα. Υπάρχουν πολλές άλλες τεχνικές, όπως οι t-closeness, δ-presence και ρ-uncertainty, όπου η κάθε μία παρέχει διαφορετικές εγγυήσεις σε σχέση με την προστασία και την ποιότητα των δεδομένων.

Οι πραγματικές προκλήσεις στις τεχνικές ανωνυμοποίησης είναι πολλές. Ενώ υπάρχουν πάρα πολλοί αλγόριθμοι ανωνυμοποίησης, έρχονται με υποθέσεις και απαιτήσεις, οι οποίες δυσκολεύουν την εφαρμογή τους. Οι τρεις κύριες απαιτήσεις για να εφαρμοστεί η ανωνυμοποίηση δεδομένων είναι :

1. Η ποσοτικοποίηση και η ρύθμιση της απώλειας πληροφορίας. Η ανωνυμοποίηση δεδομένων μετατρέπει τα αρχικά δεδομένα με τέτοιο τρόπο ώστε να εξαλείψει τις ευαίσθητες πληροφορίες, ενώ ταυτόχρονα να διατηρήσει όσον το δυνατόν περισσότερες πληροφορίες που είναι πολύτιμες για την ανάλυση. Οι περισσότεροι αλγόριθμοι το καταφέρνουν αυτό γενικεύοντας τα δεδομένα, δηλαδή αντικαθιστούν τις αρχικές τιμές με καινούργιες πιο γενικευμένες π.χ. το όνομα της πόλης, αντικαθίσταται από το όνομα της χώρας, ή αφαιρώντας τελείως τις εγγραφές με αυτές τις τιμές. Οι αλγόριθμοι ανωνυμοποίησης παρέχουν στον χρήστη, διαφορετικές λύσεις για το πως μπορούν να ανωνυμοποιηθούν τα δεδομένα. Ταυτόχρονα παρέχουν στο χρήστη για κάθε λύση και μία κλίμακα απώλειας πληροφορίας, δηλαδή αναφέρουν ποιο ποσοστό πληροφοριών αφαιρέθηκαν από τα αρχικά δεδομένα. Αυτό μπορεί να βοηθήσει τον χρήστη να επιλέξει την λύση που τον ικανοποιεί περισσότερο και ταυτόχρονα να εξασφαλίσει την ιδιωτικότητα των χρηστών.
2. Η κλιμακωσιμότητα. Τα σημαντικά σημασιολογικά ζητήματα που έχουν σχέση με τη διαφύλαξη της ιδιωτικής ζωής έχουν επισκιάσει την δυνατότητα κλιμάκωσης. Πολλές μέθοδοι ανωνυμοποίησης βασίζονται στη συσταδοποιήση των δεδομένων

ή σε άλλες δαπανηρές υπολογιστικές μεθόδους που τις καθιστούν κατάλληλες μόνο για μικρά δεδομένα.

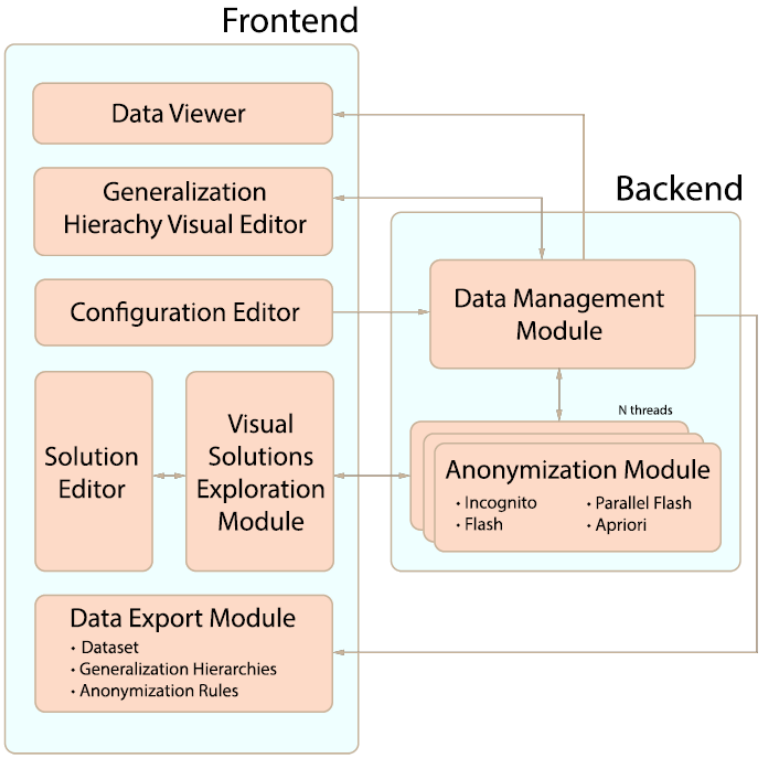
3. Η ύπαρξη συμπληρωματικού υλικού, όπως οι ιεραρχίες.. Πολλές μέθοδοι ανωνυμοποίησης βασίζονται σε ιεραρχίες ανωνυμοποίησης, δηλαδή μία ιεραρχία που περιγράφει τον τρόπο με τον οποίο μπορεί να κάθε αρχική τιμή να αντικατασταθεί από άλλες πιο αφηρημένες τιμές. Για παράδειγμα μια ιεραρχία μπορεί να προσδιορίσει ότι η Αθήνα μπορεί να γενικευτεί στην Ελλάδα. Και η Ελλάδα στην Ευρώπη. Τέτοιες πληροφορίες είναι σπάνια διαθέσιμες και εξαρτάται επίσης από τους τομείς εφαρμογής.

Η διαδικτυακή πλατφόρμα του Citysense έχει στόχο να συνδυάζει πληροφορίες που αφορούν την ανθρώπινη δραστηριότητα με χωρική πληροφορία. Τα δεδομένα που διαχειρίζεται σχετίζονται με την οικονομία, την εγκληματικότητα, διάφορες κοινωνικές δραστηριότητες κτλ., και μπορεί να αποκαλύπτουν ιδιαίτερα ευαίσθητες προσωπικές πληροφορίες. Η δημοσίευση τέτοιων δεδομένων είναι ιδιαίτερα πολύτιμη, αλλά θα πρέπει να γίνεται με τρόπο που δεν παραβιάζει την ιδιωτικότητα των χρηστών. Στο πλαίσιο του Citysense επεκτείναμε το εργαλείο Amnesia ώστε να χειρίζεται ευαίσθητη πληροφορία με χωρική αναφορά. Το Amnesia εστιάζει στην εφαρμοσιμότητα και στην φιλικότητα προς τον χρήστη και παρέχει έναν εύκολο τρόπο για την χρησιμοποίηση, την επεξεργασία και την αυτόματη παραγωγή ιεραρχιών. Ακόμη, παρέχει αποδοτικούς και κλιμακωτούς αλγόριθμους ανωνυμοποίησης για σχεσιακά και transactional δεδομένα. Οι χρήστες μπορούν να ρυθμίσουν την απώλεια πληροφοριών και την καθοδήγηση της διαδικασίας της ανωνυμοποίησης μέσω της γραφικής διερεύνησης των υποψηφίων λύσεων, παρατηρώντας στατιστικά τεχνικής εξόρυξης και την ανάμειξη των επιλογών των χρηστών για αφαίρεση εγγραφών σε λύσεις βασισμένες σε γενικεύσεις που παράγονται από αλγόριθμο.

6.6.1 Αρχιτεκτονική Συστήματος

Το εργαλείο ανωνυμοποίησης Amnesia είναι υλοποιημένο σαν web-based και σαν desktop εφαρμογή. Επιλέξαμε web τεχνολογίες για την ανάπτυξη του, καθώς οι τεχνολογίες web είναι πιο εξελιγμένες και αναπτύσσονται και πιο γρήγορα σε σχέση με τις τεχνολογίες που είναι για standalone εφαρμογές. Το Amnesia (και στις δύο εκδόσεις του) χωρίζεται σε δύο υποσυστήματα, το Frontend και το Backend, τα οποία χρησιμοποιούν διαφορετικές τεχνολογίες. Τα δύο υποσυστήματα είναι συνδεδεμένα με ένα μοντέλο client-server και η αλληλεπίδραση τους γίνεται μέσω structured web services.
Τα στοιχεία του Frontend (client) υλοποιούν το γραφικό περιβάλλον του Amnesia, το οποίο οδηγεί τους χρήστες μέσω διάφορων βημάτων στην διαδικασία ανωνυμοποίησης. Το Frontend είναι μια web-based εφαρμογή, η οποία έχει υλοποιηθεί χρησιμοποιώντας την γλώσσα προγραμματισμού JavaScript σε συνεργασία με το εργαλείο Bootstrap. Το Backend (server) είναι υλοποιημένο στην γλώσσα προγραμματισμού Java, σε συνεργασία με το εργαλείο Spring, το οποίο παρέχει όλη την λειτουργικότητα για τα web services. Το Frontend χρησιμοποιεί jQuery για την επικοινωνία με το Backend και όλα τα μηνύματα-απαντήσεις είναι σε μορφή json.
Οι παραπάνω τεχνολογίες αποτελούν τις πιο σύγχρονες τεχνολογίες για web εφαρμογές και προφέρουν πολύ σημαντικά πλεονεκτήματα: α) Είναι πολύ εύκολες στην εγκατάσταση τους σε πολλές διαφορετικές πλατφόρμες και μπορούν να συνδυαστούν με μία μεγάλη ποικιλία άλλων τεχνολογιών, β) παρέχουν ένα μεγάλο εύρος λειτουργιών και γ) είναι συμβατές με όλους τους browsers.
Θα πρέπει επίσης να αναφέρουμε κάτι σημαντικό για την ασφάλεια των δεδομένων των χρηστών, ότι τα δεδομένα δεν αποθηκεύονται εκεί που τρέχει το backend αλλά τοπικά

στον υπολογιστή μας. Μία εικονική απεικόνιση της αρχιτεκτονικής του συστήματος παρουσιάζεται στο παρακάτω σχήμα:



6.6.1.1 Frontend

Το Frontend έχει υλοποιηθεί χρησιμοποιώντας το εργαλείο Bootstrap. Το Bootstrap είναι ένα ολοκληρωμένο HTML,CSS, JavaScript εργαλείο για την υλοποίηση web γραφικών εφαρμογών, το οποίο χρησιμοποιεί μοντέρνες web πλατφόρμες για βελτιώσει και να ικανοποιήσει τις ανάγκες των χρηστών. Επίσης, έχει πολλές πρόσθετες βιβλιοθήκες και εργαλεία, τα οποία μπορούν να συνδυαστούν εύκολα και να δημιουργήσουν ένα πιο όμορφο, εύχρηστο και λειτουργικό γραφικό περιβάλλον για τους χρήστες. Το Bootstrap είναι μία πλατφόρμα ανάπτυξης και παρέχει συχνή ενημέρωση των βιβλιοθηκών και των εργαλείων που χρησιμοποιεί, έτσι ώστε να σου παρέχει πάντα τις τελευταίες τους εκδόσεις. Παρακάτω θα αναλύσουμε τις κύριες λειτουργίες της εφαρμογής και τις βασικές βιβλιοθήκες που χρησιμοποιούνται:

Οι Frontend λειτουργίες του Amnesia:

- Φόρτωση ανωνυμοποιημένων και μη δεδομένων, ιεραρχίων και κανόνων ανωνυμοποίησης.
- Αποθήκευση ανωνυμοποιημένων και μη δεδομένων, ιεραρχίων και κανόνων ανωνυμοποίησης.
- Αυτόματη δημιουργία και επεξεργασία ιεραρχιών γενίκευσης.
- Εκτέλεση αλγορίθμου ανωνυμοποίησης και γραφική παρουσίαση του αποτελέσματος του. Επιτρέπει επίσης και την ανάλυση και επεξεργασίας του αποτελέσματος.
- Ανάλυση της ποιότητας των ανωνυμοποιημένων δεδομένων. Η ανάλυση γίνεται με γραφικό τρόπο, αλλά και με ad hoc ερωτήματα πάνω στα δεδομένα.
- Δίνει την δυνατότητα στο χρήστη να διαγράψει σύνολο δεδομένων είτε στα αρχικά είτε στα τελικά δεδομένα, ανάλογα με την μετέπειτα χρήση που θέλει να τους κάνει.

- Ταυτόχρονη παράθεση των αρχικών και των τελικών δεδομένων, έτσι ώστε ο χρήστης να μπορεί εύκολα να αντιληφθεί τις αλλαγές και να αξιολόγηση το αποτέλεσμα.
- Η ικανότητα να παράγει και να εφαρμόζει κανόνες ανωνυμοποίησεις.

6.6.1.2 Γραφική απεικόνιση της Ιεραρχίας Γενίκευσης

Το Amnesia παρέχει έναν γραφικό περιβάλλον για την επεξεργασία των ιεραρχιών γενίκευσης. Το γραφικό περιβάλλον απεικονίζει τις ιεραρχίες σε μία δενδρική μορφή και επιτρέπει στους χρήστες με εύκολο τρόπο να τις επεξεργαστούν τους κόμβους του δέντρου, χρησιμοποιώντας τις λειτουργίες : προσθήκη, διαγραφή και επεξεργασία. Μία ιεραρχία μπορεί να δημιουργηθεί αυτόματα από το εργαλείο, χρησιμοποιώντας σαν βάση τις τιμές ενός συγκεκριμένου χαρακτηριστικού (μίας στήλης του πίνακα) και στην συνέχεια μπορεί ο χρήστης να την επεξεργαστεί. Οι ιεραρχίες μπορούν εύκολα να αποθηκευτούν και να φορτωθούν. Ο αριθμός εμφάνισης κάθε τις κάθε τιμής στο σύνολο δεδομένων οπτικοποιείται με διαγράμματα, για να παρέχει την κατανομή τους. Ο αριθμός αυτός για μία γενικευμένη τιμή g υπολογίζεται ως το άθροισμα όλων των αριθμών εμφάνισης των αρχικών τιμών που γενικεύονται στην τιμή g.

6.6.1.3 Εκτέλεση αλγορίθμου

Το εργαλείο επιτρέπει στον χρήστη να επιλέξει τον αλγόριθμο ανωνυμοποίησης που επιθυμεί, καθώς και να εισάγει τις αντίστοιχες παραμέτρους που χρειάζεται για την εκτέλεση του π.χ. ο αλγόριθμος *k*-anonymity χρειάζεται την παράμετρο k, ενώ ο αλγόριθμος k^m-anonymity χρειάζεται τις παραμέτρους *k* ,*m*. Στην συνέχεια το γραφικό περιβάλλον θα στείλει το αίτημα στο backend.

6.6.1.4 Γραφική απεικόνιση για την εξερεύνηση των ανωνυμοποιημένων λύσεων

Αυτή η λειτουργία οπτικοποιεί τον χώρο των λύσεων του αλγορίθμου k-anonymity σαν ένα γράφο γενίκευσης όπου κάθε κόμβος αποτελεί μία λύση. Οι κόμβοι που αντιστοιχούν σε μία λύση που εξασφαλίζει την επιθυμητή ανωνυμοποίηση απεικονίζονται με πράσινο χρώμα, ενώ οι υπόλοιποι με κόκκινο. Ο χρήστες μπορούν με γραφικό τρόπο να εξερευνήσουν τον γράφο (ζουμ) και να επιλέξουν το κόμβο με την λύση γενίκευσης που επιθυμούν (είτε οδηγεί στην επιθυμητή ανωνυμοποίηση είτε όχι) και να εφαρμόσει αυτή την λύση στα αρχικά δεδομένα.

6.6.1.5 Γραφική απεικόνιση του συνόλου των δεδομένων

Η γραφική απεικόνιση παρουσιάζει τα αρχικά και τα ανωνυμοποιημένα δεδομένα σε μία σχεσιακή μορφή. Και τα δύο σύνολα δεδομένων παρουσιάζονται δίπλα δίπλα, για να επιτρέπει στον χρήστη να εξετάζει ατομικά την κάθε εγγραφή. Έχει υλοποιηθεί με την βοήθεια της JavaScript βιβλιοθήκης DataTables, η οποία παρέχει δυναμικούς πίνακες με πολλές λειτουργικότητες. Οι πιο σημαντικές είναι : ταξινόμηση με οποιοδήποτε χαρακτηριστικό, σελιδοποίηση και αναζήτηση.

6.6.1.6 Γραφική απεικόνιση της λύσης

Αυτή η λειτουργικότητα αφορά μόνο το ανωνυμοποιημένο σύνολο δεδομένων. Αρχικά ο χρήστης μπορεί να δει ένα δείγμα του ανωνυμοποιημένου συνόλου δεδομένων για αυτή την λύση, χωρίς όμως να εφαρμοστεί αυτή η λύση. Ακόμη ο χρήστης μπορεί να παρακολουθήσει με την χρήση κάποιων διαγραμμάτων την κατανομή των τιμών στο νέο σύνολο δεδομένων. Τέλος, εάν η λύση που εξετάζει ο χρήστης δεν είναι ανωνυμοποιημένη, ο χρήστης μπορεί να διαγράψει τις εγγραφές που δημιουργούν πρόβλημα.

6.6.1.7 Φόρτωμα/Αποθήκευση Αρχείων

Το Amnesia υποστηρίζει το φόρτωμα και την αποθήκευση αρχείων. Η λειτουργία του φορτώματος μπορεί να γίνει με δύο διαφορετικούς τρόπους είτε χρησιμοποιώντας ένα

σύνδεσμο html από το κεντρικό μενού, είτε χρησιμοποιώντας την JavaScript βιβλιοθήκη Dropzone, στην οποία απλώς «σέρνουμε» το αρχείο πάνω στην είκονα. Τέλος, ο χρήστης έχει την δυνατότητα να αποθηκεύει σε αρχεία τα αποτελέσματα των αλγορίθμων.

6.6.1.8 Εξαγωγή συνόλου δεδομένων

Ο χρήστης έχει την δυνατότητα, μέσα από το εργαλείο, να αποθηκεύσει πολλές πληροφορίες που δημιουργούνται κατά την διάρκεια της διαδικασίας της ανωνυμοποίησης. Με τον τρόπο αυτό ο χρήστης μπορεί να επαναχρησιμοποιήσει στοιχεία και σε άλλα σύνολα δεδομένων. Οι πληροφορίες που μπορεί να εξάγει είναι : ιεραρχίες, αρχικό και ανωνυμοποιημένο σύνολο δεδομένων και κανόνες ανωνυμοποίησης.

6.6.1.9 Στατιστικά για την ποιότητα των δεδομένων

Μέσω του εργαλείου δίνεται στον χρήστη η δυνατότητα να αξιολογήσει την ποιότητα των δεδομένων, χρησιμοποιώντας ερωτήματα στα αρχικά και στα ανωνυμοποιημένα δεδομένα. Αναλυτικότερα ο χρήστης μπορεί να επιλέξει τα χαρακτηριστικά που τον ενδιαφέρουν και να εισάγει αντίστοιχα τις τιμές τους. Στην συνέχεια, το Amnesia θα του επιτρέψει 4 διαγράμματα :

- Τον αριθμό των εγγραφών που ικανοποιούν το ερώτημα στο αρχικό σύνολο δεδομένων
- Τον μέγιστο αριθμό εγγραφών που μπορεί να ικανοποιούν το ερώτημα στα ανωνυμοποιημένα δεδομένα. Σε αυτή την περίπτωση, προσμετράται κάθε γενίκευση των τιμών που μπορεί να έχει προέλθει από τις τιμές που υπάρχουν στο ερώτημα,.
- Τον ελάχιστο αριθμό εγγραφών που ικανοποιούν το ερώτημα στα ανωνυμοποιημένα δεδομένα. Αυτά που οι εγγραφές περιέχουν τις τιμές του ερωτήματος ή περιέχουν πιο συγκεκριμένες τιμές, εάν οι τιμές των ερωτημάτων είναι ήδη γενικευμένες.
- Μία εκτίμηση των έγγραφών που ικανοποιούν το ερώτημα στα ανωνυμοποιημένα δεδομένα.

Όταν ο χρήστης επιλέξει πάνω από ένα χαρακτηριστικό, τότε οι παραπάνω αριθμοί υπολογίζονται χρησιμοποιώντας την τομή των χαρακτηριστικών.

6.6.1.10 Επαναχρησιμοποίηση ιεραρχιών

Οι ιεραρχίες καθορίζουν κανόνες γενίκευσης για ένα σύνολο δεδομένων, αλλά μπορούν να επαναχρησιμοποιηθούν σε χαρακτηριστικά με παρόμοια πεδία ορισμού. Πιο αναλυτικά, εάν ο χρήστης συνδέσει μία ιεραρχία με ένα χαρακτηριστικό, μπορεί ταυτόχρονα να την χρησιμοποιήσει και για κάποιο άλλο στο ίδιο σύνολο δεδομένων. Η μόνη προϋπόθεση που υπάρχει για να συμβεί αυτό, είναι ότι το χαρακτηριστικό στο οποίο θα γίνει η επαναχρησιμοποίηση, θα πρέπει να έχει το ίδιο ή κάποιο υποσύνολο του εύρους τιμών του πρώτου χαρακτηριστικού.

6.6.2 Backend

Το backend του Amnesia αποτελεί ουσιαστικά την πλευρά του server και επικοινωνεί με το frontend χρησιμοποιώντας RESTful web services. Για να δημιουργήσουμε RESTful web services χρησιμοποιήσαμε το εργαλείο Spring, το οποίο παρέχει ένα ολοκληρωμένο μοντέλο προγραμματισμού και διαμόρφωσης για σύγχρονες εφαρμογές βασισμένες σε Java και για κάθε είδους πλατφόρμα ανάπτυξης. Ένα βασικό στοιχείο του Spring είναι η υποστήριξη των υποδομών σε επίπεδο εφαρμογών. Το Spring επικεντρώνεται στην «εγκατάσταση» των εφαρμογών έτσι ώστε οι ομάδες να μπορούν να επικεντρωθούν στην

λογική σε επίπεδο εφαρμογών, χωρίς περιττά προβλήματα με συγκεκριμένα περιβάλλοντα ανάπτυξης.

6.6.2.1 Διαχείριση δεδομένων

Αυτό το δομικό στοιχείο είναι υπεύθυνο για την δημιουργία των δομών δεδομένων, που απαιτούν οι αλγόριθμοι ανωνυμοποίησης. Όλες οι δομές δεδομένων αποθηκεύονται στην μνήμη. Το σύνολο δεδομένων αναπαρίσταται σαν ένα δισδιάστατο πίνακα και κάθε τιμή των δεδομένων αντιπροσωπεύεται από έναν αριθμό με την βοήθεια λεξικών. Μία γενικευμένη ιεραρχία υλοποιείται σαν ένα Java map με κλειδί τον γονικό κόμβο και τιμή έναν πίνακα διαδοχικούς κόμβους.

6.6.2.2 Αυτόματη δημιουργία ιεραρχιών

Το Amnesiaχρησιμοποιεί το αρχικό σύνολο δεδομένων και τις προτιμήσεις (μέγεθος ιεραρχίας, αριθμός παιδιών, κτλ.) του χρήστη για να δημιουργήσει μία ιεραρχία γενίκευσης. Η γενική ιδέα είναι ότι ταξινομεί όλες τις τιμές που υπάρχουν στο σύνολο δεδομένων και μετά τις χωρίζει σε ομάδες χρησιμοποιώντας τις προτιμήσεις του χρήστη και μια ομοιόμορφη κατανομή. Ο αλγόριθμος μπορεί να επεξεργαστεί και τις κενές τιμές.

6.6.2.3 Έλεγχος Ανωνυμοποίησης των δεδομένων

Με αυτή την λειτουργία ο χρήστης έχει την δυνατότητα να ελέγξει εάν το σύνολο των δεδομένων που φόρτωσε είναι ανωνυμοποιημένο ή όχι, χωρίς να χρειαστεί να τρέξει κανένα αλγόριθμο. Ο χρήστης απλώς θα πρέπει να επιλέξει τα χαρακτηριστικά που θέλει και τον αριθμό k. Στην συνέχεια το πρόγραμμά μας θα του κάνει μία αναπαράσταση της κατανομής όλων των τιμών, χρησιμοποιώντας διαγράμματα, και θα τον ενημερώσει εάν το σύνολο δεδομένων είναι ανωνυμοποιημένο ή όχι. Στην δεύτερη περίπτωση, ο χρήστης μπορεί να διαγράψει τις εγγραφές που του δημιουργούν πρόβλημα και να ανωνυμοποιήσει τα δεδομένα του, χωρίς να χρησιμοποιήσει κανέναν αλγόριθμο.

6.6.2.4 Ανωνυμοποίηση Δεδομένων

Η μεγαλύτερη και πιο σύνθετη πρόκληση στην λειτουργία του εργαλείου Amnesia είναι οι βασικοί αλγόριθμοι ανωνυμοποίησης που χρησιμοποιούνται στο backend. Το Amnesia διαθέτει αλγορίθμους για διαφορετικές μορφές δεδομένων, όπως τον παράλληλο αλγόριθμο Flash που χρησιμοποιείται σε δεδομένα που έχουν σχεσιακή μορφή και τον αλγόριθμο Apriori για δεδομένα που έχουν set-valued μορφή.

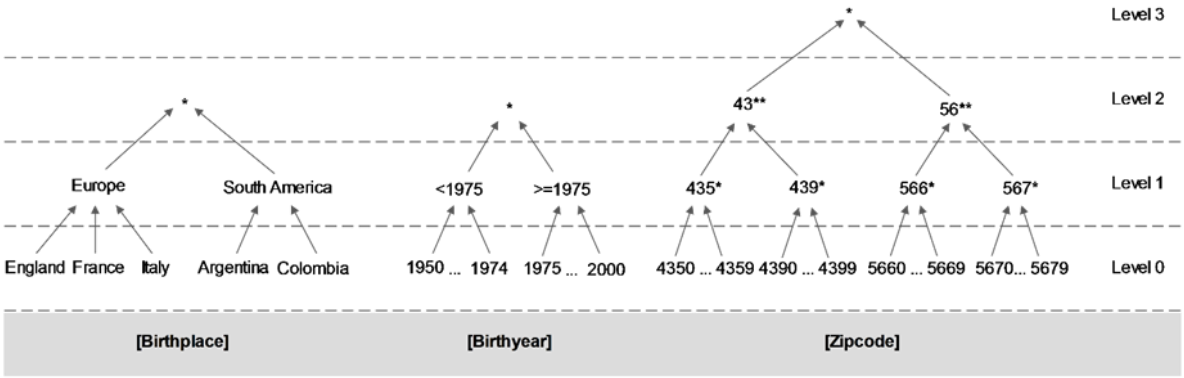
Και οι δύο αλγόριθμοι χρησιμοποιούν γενικεύσεις για να μετατρέψουν τα δεδομένα σε ένα ανωνυμοποιημένο σύνολο δεδομένων. Η γενίκευση είναι η αντικατάσταση μίας τιμής με μία πιο γενικευμένη τιμή. Αυτή η αντικατάσταση γίνεται σύμφωνα με την ιεραρχία γενίκευσης, η οποία ομαδοποιεί παρόμοιες τιμές των αρχικών δεδομένων και καθορίζει τον τρόπο με τον οποίο θα αντικατασταθούν από τον αλγόριθμο. Δίνεται ένα παράδειγμα για τρεις διαφορετικές ιεραρχίες γενίκευσης, τόπος γέννησης (Birthplace), έτος γέννησης(Birthdate) και ταχυδρομικούς κώδικες(Zipcode). Η ιδέα είναι ότι οι όροι που είναι σημασιολογικά κοντά πχ η Γαλλία και η Ιταλία σύμφωνα με την τοποθεσία μπορούν να αντικατασταθούν με κάποιο πιο γενικευμένο όρο, όπως είναι η Ευρώπη. Οι αλγόριθμοι ανωνυμοποίησης χρησιμοποιούν αυτές τις ιεραρχίες για να μετατρέψουν τα δεδομένα σε ανωνυμοποιημένα δεδομένα, έτσι ώστε να τηρούν την επιθυμητή ασφάλεια. Υπάρχουν πολλοί τρόποι για να χρησιμοποιήσουμε μία ιεραρχία γενίκευσης. Οι αλγόριθμοί χρησιμοποιούν τα παρακάτω:

- Global recording. Ο αλγόριθμος εκτελεί αντικατάσταση όλων των ίδιων τιμών που βρίσκονται στα αρχικά δεδομένα με μία γενίκευση τους. Π.χ. η Γαλλία γενικεύεται στην Ευρώπη, οπότε κάθε εμφάνιση της Γαλλίας στα αρχικά δεδομένα θα

αντικαθίσταται με την τιμή Ευρώπη. Στο local recording, αυτή η αντικατάσταση είναι μερική.

- Full-domain generalizations. Η γενίκευση πλήρους πεδίου υποδηλώνει ότι εάν ο αλγόριθμος γενικεύει μια μοναδική τιμή στο επόμενο επίπεδο αφαίρεσης, τότε κάθε τιμή θα γενικευθεί στο ίδιο επίπεδο αφαίρεσης. Για παράδειγμα, αν ο αλγόριθμος αποφασίσει ότι πρέπει να γενικεύσει τη Γαλλία στην Ευρώπη, τότε κάθε χώρα θα γενικευτεί στην ήπειρο, π.χ. η Αργεντινή θα πρέπει να γενικευτεί στη Νότια Αμερική. Όταν χρησιμοποιείται η γενίκευση πλήρους πεδίου, τότε όλες οι τιμές γενικεύονται πάντοτε στο ίδιο επίπεδο γενίκευσης στην αντίστοιχη ιεραρχία. Αυτή η γενίκευση χρησιμοποιείται για τα σχεσιακά δεδομένα στον αλγόριθμο Flash του Amnesia.

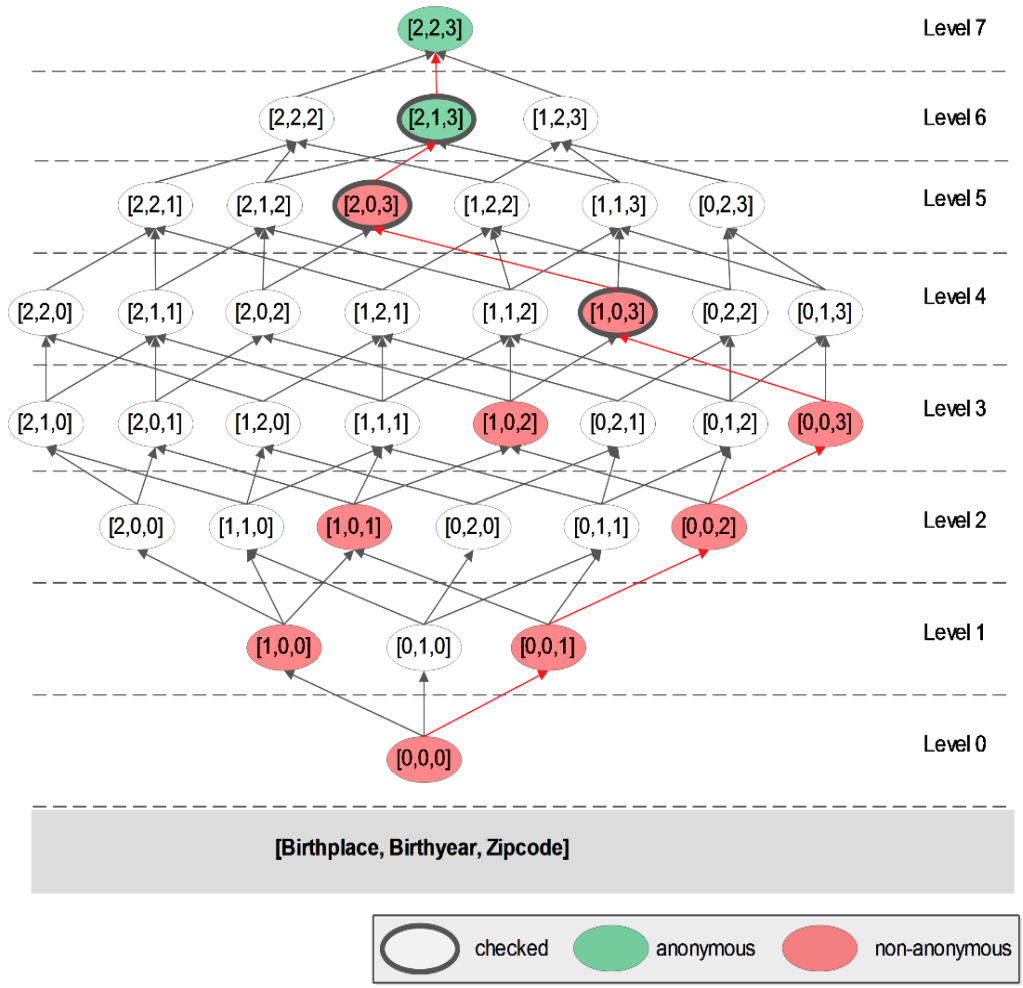
- Partial-domain generalizations. Σε αντίθεση με τη γενίκευση πλήρους πεδίου, η μερική γενίκευση πεδίου, δεν απαιτεί τη γενίκευση κάθε τιμής του ίδιου χαρακτηριστικού στο ίδιο επίπεδο αφαίρεσης. Για παράδειγμα, οι ευρωπαϊκές χώρες μπορούν να γενικευτούν στην Ευρώπη, αλλά οι χώρες της Νότιας Αμερικής μπορούν να εμφανιστούν αμετάβλητες εάν διατηρείται η εγγύηση προστασίας της ιδιωτικότητας. Σημειώστε ότι ακόμα και σε αυτή την περίπτωση όλοι οι αδελφοί κόμβοι μιας τιμής είναι πάντα γενικευμένοι μαζί, δηλαδή αν η Γαλλία γενικευτεί στην Ευρώπη, τότε η Αγγλία και η Ιταλία θα πρέπει επίσης να γενικευτούν στην Ευρώπη.



6.6.2.5 Flash Algorithm

Ο αλγόριθμος FLash είναι ένας εξαντλητικός αλγόριθμος που βρίσκει την καλύτερη k-ανώνυμη λύση για σχεσιακά δεδομένα, χρησιμοποιώντας το πλήρες μοντέλο γενίκευσης πεδίου. Όπως και ο αλγόριθμος Incognito, δημιουργεί το πλήρες πλέγμα γενίκευσης και στη συνέχεια ψάχνει για την καλύτερη λύση. Το πλέγμα γενίκευσης, αντιπροσωπεύει ολόκληρο το χώρο λύσης για ένα δεδομένο σύνολο δεδομένων και ιεραρχιών. Κάθε κόμβος αντιπροσωπεύει μια διαφορετική διαμόρφωση γενίκευσης, δηλαδή διαφορετικά επίπεδα γενίκευσης για κάθε χαρακτηριστικό. Στην γενίκευση πλήρους πεδίου, οι τιμές κάθε χαρακτηριστικού γενικεύονται σε τιμές που ανήκουν σε ένα επίπεδο στην ιεραρχία γενίκευσης. Ακόμα, δεν είναι όλα τα χαρακτηριστικά γενικευμένα στο ίδιο επίπεδο. Για παράδειγμα, εάν έχουμε τρία χαρακτηριστικά: τον τόπο γέννησης, το έτος γέννησης και τον ταχυδρομικό κώδικα για φυσικά πρόσωπα, τα οποία αντιστοιχούν στις αντίστοιχες ιεραρχίες, μπορούμε να έχουμε μια λύση όπου οι χώρες γενικεύονται σε ηπείρους (Επίπεδο 1 στην πρώτη ιεραρχία), οι ημερομηνίες γέννησης δεν έχουν γενικευτεί (επίπεδο 0 στη δεύτερη ιεραρχία) και οι ταχυδρομικοί κώδικες έχουν καταργηθεί (επίπεδο 1) στην τρίτη και τελευταία ιεραρχία.

Αυτή η λύση αντιπροσωπεύει έναν κόμβο στο πλέγμα γενίκευσης, το οποίο σημειώνουμε με το [1,0,1], δηλώνοντας έτσι το επίπεδο γενίκευσης κάθε χαρακτηριστικού. Το πλέγμα διατάσσεται όπως στο σχήμα 12(αλλαγή), όπου οι κόμβοι κάθε οριζόντιου επιπέδου έχουν τον ίδιο αριθμό συνολικών επιπέδων γενίκευσης. Ο αλγόριθμος Flash χρησιμοποιεί μια κατά βάθος στρατηγική, η οποία του επιτρέπει να απορρίπτει διάφορες λύσεις και να αποφεύγει να τις εξετάζει. Για κάθε κόμβο σε κάθε επίπεδο, αν ο κόμβος δεν έχει ήδη επισημανθεί (ανώνυμος ή ανώνυμος), ο Flash δημιουργεί μια διαδρομή προς τον κορυφαίο κόμβο, εφαρμόζοντας μια άπληστη στρατηγική κατά βάθος. Η δημιουργία μιας διαδρομής βασίζεται σε μια κατακόρυφη στρατηγική μετατόπισης που στοχεύει στην επιλογή κόμβων σύμφωνα με τρία σταθερά κριτήρια: (α) το συνολικό επίπεδο γενίκευσης του κόμβου στο πλέγμα, (β) τη μέση γενίκευση όλων των ψεύδοαναγνωριστικων του κόμβου Και (γ) το μέσο όρο του αριθμού διακριτών τιμών στο τρέχον επίπεδο κάθε ψεύδοαναγνωριστικου. Η αναζήτηση τερματίζεται όταν φτάνει στον κορυφαίο κόμβο ή όταν ο τρέχων κόμβος δεν



έχει διάδοχο που δεν έχει ήδη επισημανθεί. Όταν δημιουργηθεί μια διαδρομή, ο αλγόριθμος αρχίζει να ελέγχει την k-ανωνυμία με μια δυαδική στρατηγική αναζήτησης. Ξεκινά πρώτα με τον κόμβο στο μέσο της διαδρομής και στη συνέχεια συνεχίζει με τη διαδρομή προς το κάτω μέρος του πλέγματος ή προς τα επάνω, ανάλογα με το αν ο μεσαίος κόμβος ήταν ανώνυμος ή όχι. Κάθε φορά που ελέγχεται ένας κόμβος, προβλέπει και για άλλους κόμβους μέσα σε ολόκληρο το πλέγμα γενίκευσης, εάν είναι ανώνυμοι ή όχι. Αυτό επιτρέπει την αποφυγή της εξέτασης άλλων κόμβων. Για παράδειγμα, αν ένας κόμβος δεν είναι ανώνυμος, τότε όλοι οι κόμβοι σε όλες τις διαδρομές από το κάτω μέρος του

πλέγματος στον κόμβο δεν θα είναι ανώνυμοι. Ο αλγόριθμος συνεχίζεται έως ότου ελεγχτούν όλοι οι κόμβοι στο πλέγμα γενίκευσης για ανωνυμία.

Για παράδειγμα, το προηγούμενο σχήμα δείχνει την πρώτη επανάληψη του αλγόριθμου Flash. Μια διαδρομή κατασκευάζεται από τον κόμβο ρίζας [0,0,0] για να φτάσει στον κορυφαίο κόμβο [2,2,3]. Αυτή η διαδρομή περιέχει κόμβους που συνδέονται με κόκκινα βέλη στο σχήμα. Στη συνέχεια, ο Flash ελέγχει τον κόμβο 2-ανωνυμίας [1,0,3] ο οποίος είναι ο μεσαίος κόμβος της διαδρομής. Δεδομένου ότι αυτός ο κόμβος δεν είναι 2-ανώνυμος, όλοι οι προκάτοχοι αυτού του κόμβου είναι επισημαίνονται επίσης ως μη-ανώνυμοι και ο αλγόριθμος συνεχίζει εξετάζοντας το άνω μισό μονοπάτι που περιέχει κόμβους [1,0,3], [2,0,3], [2,1,3] και [2,2,3]. Και πάλι ο μέσος κόμβος [2,1,3] ελέγχεται, ο οποίος είναι 2-ανώνυμος, έτσι ο διάδοχος [2,1,3] είναι επίσης ανώνυμος και ο προκάτοχος [2,0,3] ελέγχεται για το εάν είναι ανώνυμος, αλλά δεν είναι. Τέλος, όλοι οι κόμβοι της διαδρομής έχουν ελεγχθεί και ο αλγόριθμος συνεχίζει με την ίδια διαδικασία για τους κόμβους του επιπέδου 1. Δύο από αυτούς έχουν επισημανθεί από το προηγούμενο στάδιο έτσι μόνο ο κόμβος [0,1,0] είναι υποψήφιος, από τον οποίο ο Flash θα κατασκευάσει μια νέα διαδρομή προς τον κορυφαίο κόμβο και θα ακολουθηθεί η προηγούμενη διαδικασία. Αυτή η διαδικασία συνεχίζεται μέχρι να ελεγχθούν όλοι οι κόμβοι του πλέγματος.

Η στρατηγική μετατόπισης που χρησιμοποιείται από τον Flash δίνει ένα σαφές πλεονέκτημα έναντι της κατά πλάτος στρατηγικής του Incognito. Στο Amnesia υιοθετούμε αυτή τη στρατηγική, και χρησιμοποιούμε παράλληλες διεργασίες στη διαδικασία ελέγχου κόμβων για ανωνυμία.

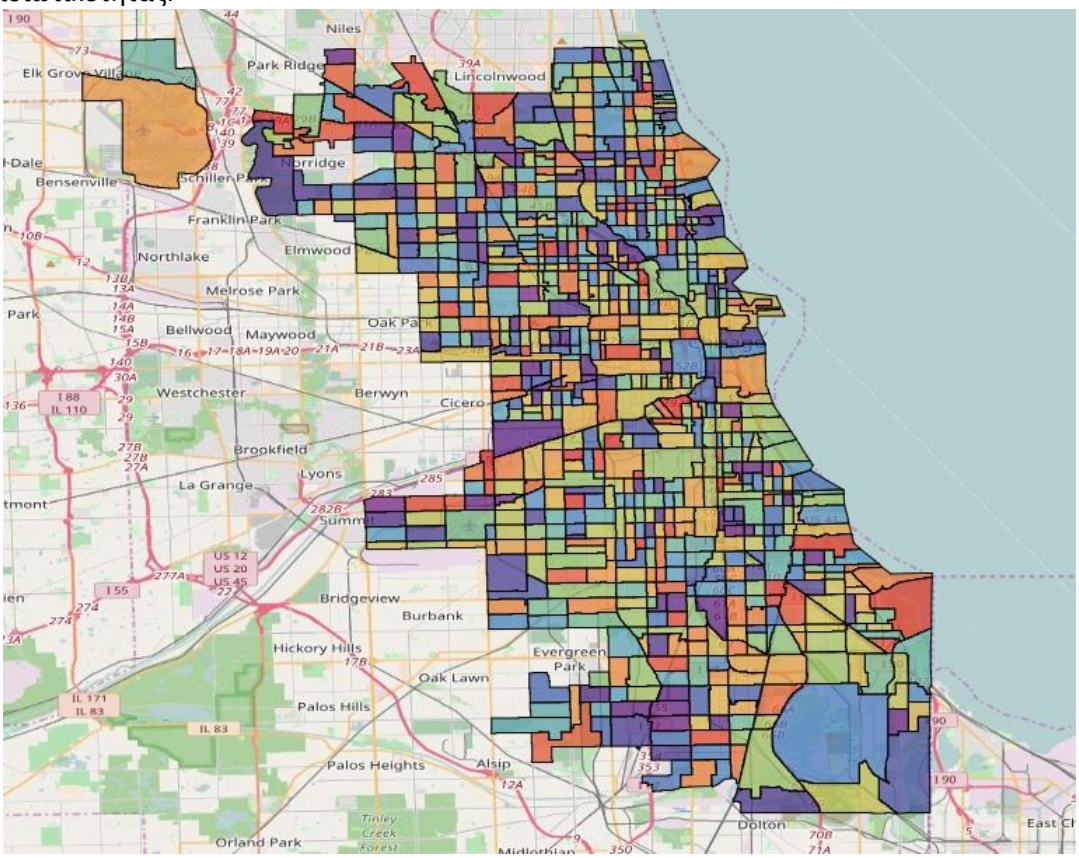
Ο έλεγχος για το εάν ένας μεμονωμένος κόμβος μετασχηματισμού είναι k-ανώνυμος ή όχι μπορεί να είναι μια χρονοβόρα εργασία. Για να ελέγξει αυτή την συνθήκη, ο αλγόριθμος πρέπει να γενικεύσει κάθε ψεύδοαναγνωριστικο στο συγκεκριμένο επίπεδο που ορίζεται από τον μετασχηματισμό και στη συνέχεια να απαριθμήσει τις σειρές που έχουν ίδιες τιμές, έτσι ώστε να καθορίσει k-ανωνυμία.

Για να επιταχυνθεί αυτή η διαδικασία, η Speedy δημιουργεί n νήματα και χωρίζει τον αρχικό πίνακα T σε n υπο-πίνακες $T_1^*, T_2^* \dots T_n^*$ με μέγεθος ίσο με |T| / n. Στη συνέχεια, κάθε νήμα i εκτελεί ξεχωριστά τη διαδικασία γενίκευσης του υπό-πίνακα T_i^* και επιστρέφει στο κύριο νήμα ένα map που περιλαμβάνει διάφορους συνδυασμούς γενικευμένων ψεύδοαναγνωριστικων που βρέθηκαν σε αυτόν τον υπό-πίνακα και τον αντίστοιχο αριθμό τους. Τέλος, για να προσδιοριστεί αν ο τρέχων μετασχηματισμός είναι k-ανώνυμος, το κύριο νήμα απλά συγχωνεύει τα αποτελέσματα και ελέγχει εάν κάθε συνδυασμός τιμών υπάρχει περισσότερο από k φορές.

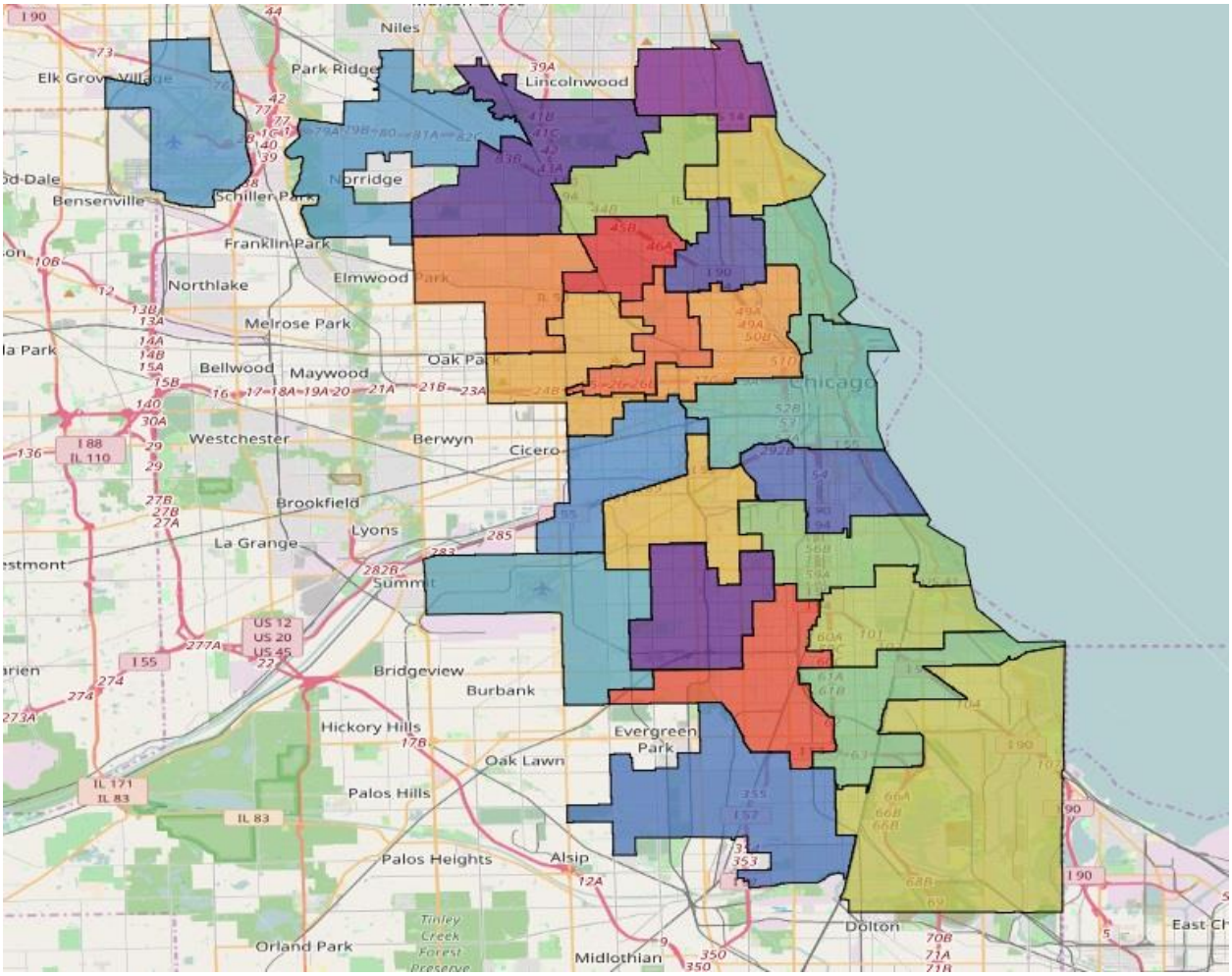
6.6.3 Ανωνυμοποίηση χωρικών δεδομένων

Το CitySense παρέχει χρήσιμες πληροφορίες στους χρήστες, από διάφορες πηγές π.χ. στοιχεία για εγκλήματα(π.χ. αριθμούς εγκλημάτων σε κάθε περιοχή) , από κοινωνικά δίκτυα (π.χ. πόσοι χρήστες, σε ποια περιοχή και τι ώρα έκαναν check-in) κα., συνδεδεμένα με χωρική πληροφορία. Εάν κάποιος κακόβουλος χρήστης χρησιμοποιήσει τις παραπάνω πληροφορίες και τις συσχετίσει με άλλα δεδομένα που μπορεί να έχει στην διάθεση του(π.χ. ποιοι κατοικούν σε κάθε περιοχή, κ.α.), μπορεί να εξάγει πληροφορίες για συγκεκριμένους χρήστες. Για αυτό το λόγο αποφασίσαμε να υλοποιήσουμε ένα νέο αλγόριθμο στο εργαλείο ανωνυμοποίησης Amnesia, που ανωνυμοποιεί γεωχωρικά δεδομένα. Για να κατανοήσουμε το πρόβλημα τις ιδιωτικότητας στην συγκεκριμένη περίπτωση θα αναλύσουμε ένα τέτοιο σενάριο. Εάν μία περιοχή έχει π.χ., έναν πολύ μικρό αριθμό εγκλημάτων, τότε ένας κακόβουλος χρήστης μπορεί εύκολο με τις πληροφορίες από το CitySense να βρει και την ακριβής ημερομηνία που έγινε. Στην περίπτωση που αυτός ο χρήστης έχει πρόσβαση σε

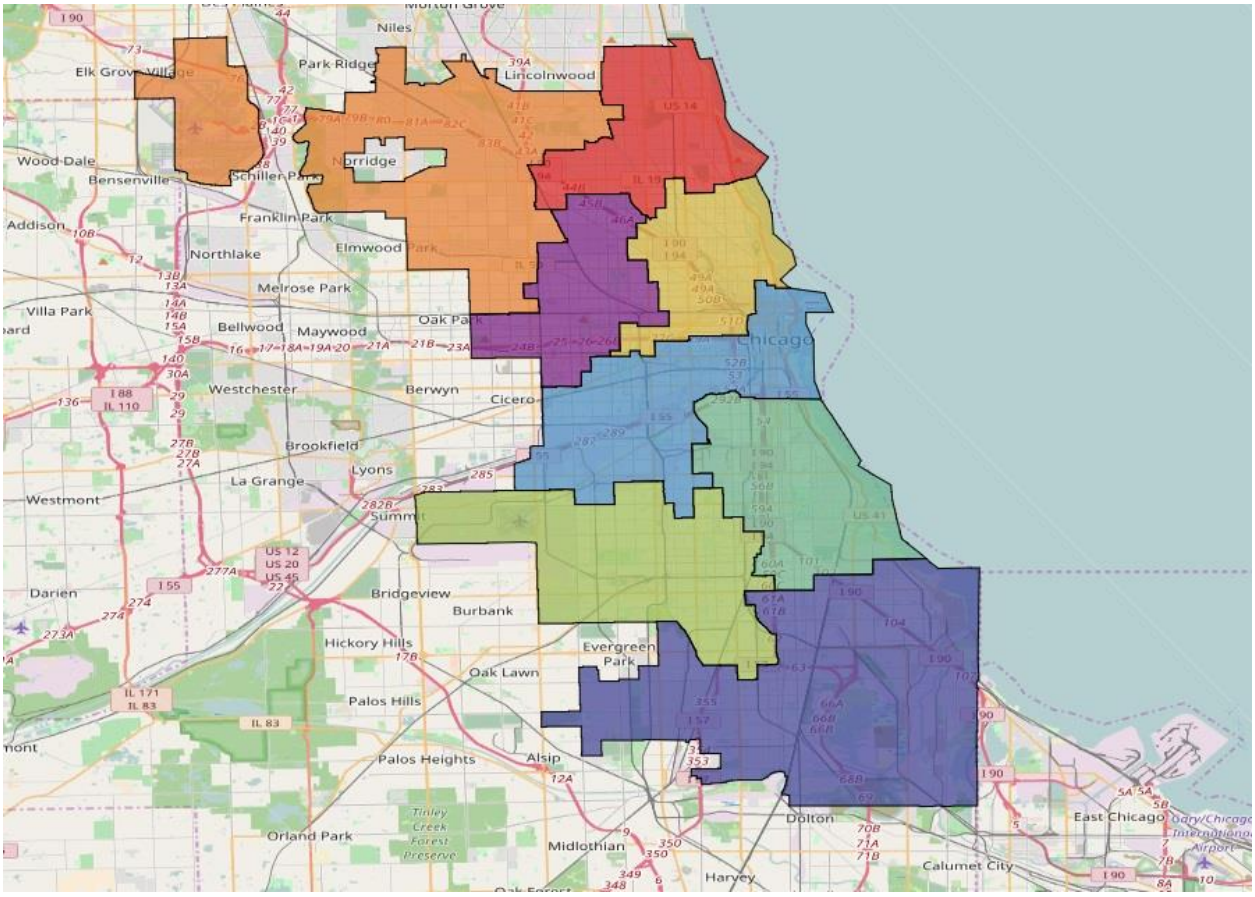
κάποιες άλλες πληροφορίες, όπως ποιοι κάτοικοι έχουν καταδικαστεί για εγκληματικές ενέργειες, τότε κάνοντας μία συσχέτιση με των πληροφοριών, μπορεί να βρει και το όνομα του και το έγκλημα που διέπραξε. Αυτό αποτελεί ένα παράδειγμα παραβίασης της ιδιωτικότητας.



Η βασική διαφορά με τα σχεσιακά δεδομένα είναι ότι σε αυτή την περίπτωση, τα ψευδοαναγνωριστικά είναι δυσδιάστατα δεδομένα, και περιγράφουν περιοχές στον χώρο. Για παράδειγμα η παραπάνω εικόνα παρουσιάζει το παράδειγμα του Σικάγο. Κάθε χρωματισμένη περιοχή είναι το ψευδοαναγνωριστικό μίας περιοχής για την οποία έχουν στοιχεία. Η εικόνα είναι μέρος ενός παραδείγματος όπου έχουμε χρησιμοποιήσει πληροφορίες για τα εγκλήματα που έχουν γίνει στις περιοχές του Σικάγο. Οι πληροφορίες που μας παρέχουν τα δεδομένα για τα εγκλήματα είναι : ακριβής τοποθεσία, και ακριβής ώρα και ημερομηνία και το έγκλημα που συνέβη. Ο συνολικός αριθμός των περιοχών που είναι χωρισμένο το Σικάγο είναι 801. Ο αλγόριθμος ανωνυμοποίησης πρέπει να γενικεύσει τα ψευδοαναγνωριστικά, δηλαδή τις περιοχές έτσι ώστε να διασφαλίζεται ότι υπάρχουν τουλάχιστον κ εγγραφές ανά περιοχή. Ταυτόχρονα, θέλουμε οι γενικεύσεις που θα γίνουνε να έχουν την ελάχιστη απώλεια πληροφορίας από το αρχικό σύνολο δεδομένων. Ο αλγόριθμος δουλεύει παρόμοια με το Flash, αλλά χειρίζεται τώρα δυσδιάστατα δεδομένα και δυσδιάστατες ιεραρχίες γενίκευσεις. Οι ιεραρχίες γενίκευσης είναι δείχνουν πως μπορούν οι αρχικές περιοχές να αντικατασταθούν από ευρύτερες περιοχές. Η αντικατάσταση γίνεται με τέτοιο τρόπο ώστε οι ευρύτερες περιοχές είναι συνενώσεις των αρχικών. Το αποτέλεσμα μίας εκτέλεσης του αλγορίθμου μας φαίνεται στο παραπάνω σχήμα. Εκτός από τις περιοχές, ως ψευδοαναγνωριστικά χρησιμοποιήσαμε τον μήνα και τον χρόνο αναφοράς. Όπως παρατηρούμε, οι περιοχές από 801, έχουν γίνει πλέον έχουν γίνει 24. Αυτό οφείλεται, στο ότι υπήρχαν λίγα εγκλήματα τον ίδιο μήνα και έτος σε πολλές περιοχές, οπότε για να προστατευθούν τα ευαίσθητα δεδομένα των χρηστών, ο αλγόριθμος έπρεπε να ανέβει στα υψηλότερα επίπεδα γενίκευσης.



Στην επόμενη εικόνα παραθέτουμε μία άλλη εκτέλεση του αλγορίθμου, για διαφορετικό σύνολο δεδομένων. Σε αυτή την εκτέλεση έχουν χρησιμοποιηθεί σαν πρόσθετα ψευδοαναγνωριστικά το έτος και οι ώρες που έχουν συμβεί τα εγκλήματα. Και σε αυτή την περίπτωση ο αλγόριθμος χρειάστηκε να ανέβει σε υψηλό επίπεδο των ιεραρχιών, για να έχουμε επιτυχημένη ανωνυμοποίηση.


Niles
Park Ridge
Lincolnwood
Elk Grove Village
Wood Dale
Bensenville
Schiller Park
Addison
Franklin Park
Northlake
Elmwood
Melrose Park
Villa Park
Bellwood
Maywood
Oak Park
Westchester
Berwyn
Cicero
Brookfield
La Grange
Lyons
Westmont
Summit
Darien
Bridgeview
Burbank
Hickory Hills
Evergreen Park
Oak Lawn
Palos Hills
Palos Heights
Alsip
Orland Park
Midlothian
Dolton
East Chicago
Calumet City
Harvey
Tinley Creek Forest Preserve
I 90
I 55
I 88
IL 110
IL 171
IL 83
US 12
US 20
US 45
US 14
US 41

7 Συμπεράσματα

Στο παρόν έγγραφο παρουσιάσαμε το CitySense, ένα δυναμικό σύστημα παρουσίασης αστικών περιοχών, το οποίο παρέχει μία πλούσια οπτικοποίηση της ζωής μιας πόλης, με την ενσωμάτωση ετερογενών συνόλων δεδομένων. Με τη βοήθεια της εφαρμογής δίνονται απαντήσεις σε ερωτήματα και αποκαλύπτονται διάφορες πλευρές της ζωής της πόλης, οι οποίες δεν θα ήταν εμφανείς με την απλή παρατήρηση των συνόλων δεδομένων. Για την επίτευξη αυτού του στόχου, αναπτύξαμε ειδικά εργαλεία συλλογής και διαχείρισης δεδομένων, πλούσιες λετουργίες οπτικοποίησης και φιλτραρίσματος, και αντιμετωπίσαμε αρκετές τεχνικές δυσκολίες. Η υλοποίηση της λειτουργίας της δυναμικής οπτικοποίησης δεδομένων κοινωνικών δικτύων (δημοσιεύσεις στο Twitter, check-in, hashtags) αποτελεί μελλοντική δουλειά. Η υποστήριξη δυναμικών συνόλων δεδομένων θα μπορούσε να χρησιμοποιηθεί για την κάλυψη δεδομένων κατανάλωσης ισχύος της πόλης και δεδομένων κίνησης. Επιπλέον μελλοντικό στόχο αποτελεί η ενσωμάτωση πληροφορίας οδικού δικτύου στο σύστημά μας. Οι χρήστες θα μπορούν να υπολογίσουν πραγματικές αποστάσεις ανάμεσα σε σημεία ενδιαφέροντος χρησιμοποιώντας ειδικές λειτουργίες του CitySense βασισμένες στο οδικό δίκτυο. Τέλος, με την αύξηση των δεδομένων που ενσωματώνονται στο CitySense, θα προκύψει το ζήτημα της δυνατότητας για κλιμάκωση (scale). Μελετάται, συνεπώς, η περίπτωση χρήσης cloud υποδομής δεδομένων για την εξυπηρέτηση των μελλοντικών αναγκών αποθήκευσης και διαχείρισης δεδομένων του CitySense.

8 Αναφορές

[1] Adomavicius, Gediminas, and Alexander Tuzhilin. "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions." Knowledge and Data Engineering, IEEE Transactions on 17.6 (2005): 734-749.

[2] Alonso, Omar, et al. "Detecting uninteresting content in text streams." SIGIR Crowdsourcing for Search Evaluation Workshop. 2010.

[3] Armentano, Marcelo G., Daniela Godoy, and Analía Amandi. "Topology-based recommendation of users in micro-blogging communities." Journal of Computer Science and Technology 27.3 (2012): 624-634.

[4] Balabanović, Marko, and Yoav Shoham. "Fab: content-based, collaborative recommendation." Communications of the ACM 40.3 (1997): 66-72.

[5] Bhattacharya, Parantapa, et al. "Inferring user interests in the twitter social network." Proceedings of the 8th ACM Conference on recommender systems. ACM, 2014.

[6] Chen, Kailong, et al. "Collaborative personalized tweet recommendation." Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012.

[7] Diaz-Aviles, Ernesto, et al. "What is happening right now... that interests me?: online topic discovery and recommendation in twitter." Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012.

[8] Duan, Yajuan, et al. "An empirical study on learning to rank of tweets." Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, 2010.

[9] Ehrlinger, Lisa and Wolfram Wöß. "Towards a Definition of Knowledge Graphs." SEMANTiCS (2016).

[10] Elmongui, Hicham G., et al. "TRUPI: Twitter Recommendation Based on Users' Personal Interests." Computational Linguistics and Intelligent Text Processing. Springer International Publishing, 2015. 272-284.

[11] Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, and others, "Building Watson: An overview of the DeepQA project," AI magazine, vol. 31, no. 3, pp. 59–79, 2010.

[12] Gabrilovich, Evgeniy, and Shaul Markovitch. "Computing semantic relatedness using wikipedia-based explicit semantic analysis." IJCAI. Vol.7.2007.

[13] Hannon, John, Mike Bennett, and Barry Smyth. "Recommending twitter users to follow using content and collaborative filtering approaches." Proceedings of the fourth ACM conference on Recommender systems. ACM, 2010.

[14] Hong, Liangjie, Aziz S. Doumith, and Brian D. Davison. "Co-factorization machines: modeling user interests and predicting individual decisions in twitter." Proceedings of the sixth ACM international conference on Web search and data mining. ACM, 2013.

[15] Hwang, Frank K., Dana S. Richards, and Pawel Winter. The Steiner tree problem. Vol. 53. Elsevier, 1992.

[16] IJntema, Wouter, et al. "Ontology-based news recommendation." Proceedings of the 2010 EDBT/ICDT Workshops. ACM, 2010.

[17] Kawamae, Noriaki. "Trend analysis model: trend consists of temporal words, topics, and timestamps." Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011.

[18] Kim, Younghoon, and Kyuseok Shim. "Twitobi: A recommendation system for twitter using probabilistic modeling." Data Mining (ICDM), 2011 IEEE 11th International Conference on. IEEE, 2011.

[19] Koren, Yehuda. "Factorization meets the neighborhood: a multifaceted collaborative filtering model." Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008.

[20] Lauw, Hady W., Alexandros Ntoulas, and Krishnaram Kenthapadi. "Estimating the quality of postings in the real-time web." Proc. of SSM conference. 2010.

[21] Liu, Yang, et al. "A User Adaptive Model for Followee Recommendation on Twitter." International Conference on Computer Processing of Oriental Languages. Springer International Publishing, 2016.

[22] Lu, Chunliang, Wai Lam, and Yingxiao Zhang. "Twitter user modeling and tweets recommendation based on wikipedia concept graph." Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence.2012.

[23] Naveed, Nasir, et al. "Bad news travel fast: A content-based analysis of interestingness on twitter." Proceedings of the 3rd International Web Science Conference. ACM, 2011.

[24] Pazzani, Michael J., Jack Muramatsu, and Daniel Billsus. "Syskill & Webert: Identifying interesting web sites." AAAI/IAAI, Vol. 1. 1996.

[25] Pennacchiotti, Marco, et al. "Making your interests follow you on twitter." Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012.

[26] Pla-Karidi Danae, Yannis Stavrakas, Yannis Vassiliou. "A Personalized Tweet Recommendation Approach Based on Concept Graphs. " The 13th IEEE International Conference on Ubiquitous Intelligence and Computing (UIC), 2016.

[27] Pla-Karidi Danae. "From user graph to Topics Graph: Towards twitter followee recommendation based on knowledge graphs." Data Engineering Workshops (ICDEW), 2016 IEEE 32nd International Conference on. IEEE, 2016.

[28] Ramage, Daniel, Susan T. Dumais, and Daniel J. Liebling. "Characterizing Microblogs with Topic Models." ICWSM 10 (2010): 1-1.

[29] Rendle, Steffen, and Lars Schmidt-Thieme. "Online-updating regularized kernel matrix factorization models for large-scale recommender systems." Proceedings of the 2008 ACM conference on Recommender systems. ACM, 2008.

[30] Rendle, Steffen. "Factorization machines with libFM." ACM Transactions on Intelligent Systems and Technology (TIST) 3.3 (2012): 57.

[31] Rodríguez, Fernando M., Luis M. Torres, and Sara E. Garza. "Followee recommendation in Twitter using fuzzy link prediction." Expert Systems 33.4 (2016): 349-361.

[32] Romero, Daniel M., et al. "Influence and passivity in social media." Machine learning and knowledge discovery in databases. Springer Berlin Heidelberg, 2011. 18-33.

[33] Sarwar, Badrul, et al. "Item-based collaborative filtering recommendation algorithms." Proceedings of the 10th international conference on World Wide Web. ACM, 2001.

[34] Sarwar, Badrul M., et al. "Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering." Proceedings of the fifth international conference on computer and information technology. Vol. 1. 2002.

[35] Schafer, J. Ben, et al. "Collaborative filtering recommender systems." The adaptive web. Springer Berlin Heidelberg, 2007. 291-324.

[36] Schein, Andrew I., et al. "Methods and metrics for cold-start recommendations." Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2002.

[37] Sharma, Aneesh, et al. "GraphJet: real-time content recommendations at twitter." Proceedings of the VLDB Endowment 9.13 (2016): 1281-1292.

[38] Shi, Yue, Martha Larson, and Alan Hanjalic. "Exploiting user similarity based on rated-item pools for improved user-based collaborative filtering." Proceedings of the third ACM conference on Recommender Systems. ACM, 2009.

[39] Su, Xiaoyuan, and Taghi M. Khoshgoftaar. "A survey of collaborative filtering techniques." Advances in artificial intelligence 2009 (2009): 4.

[40] Tajbakhsh, Mir Saman, and Jamshid Bagherzadeh. "Microblogging Hash Tag Recommendation System Based on Semantic TF-IDF: Twitter Use Case." Future Internet of Things and Cloud Workshops (FiCloudW), IEEE International Conference on. IEEE, 2016.

[41] Uysal, Ibrahim, and W. Bruce Croft. "User oriented tweet ranking: a filtering approach to microblogs." Proceedings of the 20th ACM international conference on Information and knowledge management. ACM,2011.

[42] Wang, Xuerui, and Andrew McCallum. "Topics over time: a non-Markov continuous-time model of topical trends." Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining.ACM,2006.

[43] Weng, Jianshu, et al. "Twitterrank: finding topic-sensitive influential twitterers." Proceedings of the third ACM international conference on Web search and data mining. ACM, 2010.

[44] Yang, Min-Chul, and Hae-Chang Rim. "Identifying interesting Twitter contents using topical analysis." Expert Systems with Applications 41.9 (2014):4330-4336.

[45] Yang, Min-Chul, et al. "Finding interesting posts in twitter based on retweet graph analysis." Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012.

[46] Yigit, Melike, Bilal E. Bilgin, and Adem Karahoca. "Extended topology based recommendation system for unidirectional social networks." Expert Systems with Applications 42.7 (2015): 3653-3661.

[47] Bing Liu. 2012. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

[48] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In Proceedings of EMNLP, pages 79–86, Philadelphia, Pennsylvania, USA.

[49] Peter Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of ACL, pages 417–424, Philadelphia, Pennsylvania, USA.

[50] Ivan Titov and Ryan T. McDonald. 2008. Modeling online reviews with multi-grain topic models. In WWW, pages 111–120, New York, NY, USA. ACM.

[51] Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In Proceedings of the 12th IEEE International Conference on Data Mining, ICDM '12, pages 1020–1025, Brussels, Belgium.

[52] Lei Zhang and Bing Liu. 2014. Aspect and Entity Extraction for Opinion Mining", book chapter in Data Mining and Knowledge Discovery for Big Data: Methodologies, Challenges, and Opportunities, Springer, 2014.

[53] Gayatree Ganu, Noemie Elhadad, and Amelie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In Proceedings of WebDB, Providence, Rhode Island, USA.

[54] Minqing Hu and Bing Liu. 2004a. Mining and summarizing customer reviews. In Proceedings of KDD, pages 168–177, Seattle, WA, USA.

[55] Bing Liu. 2010. Sentiment analysis and subjectivity. Handbook of natural language processing 2, 627-666.

[56] Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In Proceedings of the 9th International Workshop on Semantic Evaluation, Denver, Colorado.

[57] Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeny Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. (2016). SemEval-2016 Task 5: Aspect Based Sentiment Analysis. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), June 16-17, San Diego, California. ©Association For Computational Linguistics.

[58] Tun Thura Thet, Jin-Cheon Na, and Christopher S. G. Khoo. 2010. Aspect-based sentiment analysis of movie reviews on discussion boards. J. Information Science, 36(6):823–848.

[59] Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In Proceedings of the 18th ACM Conference on Information and Knowledge Management, pages 375–384.

[60] Qiaozhu Mei , Xu Ling , Matthew Wondra , Hang Su , and ChengXiang Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs, Proceedings of the 16th international conference on World Wide Web, May 08-12, 2007, Banff, Alberta, Canada

[61] Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In Proceedings of NAACL, pages 804–812, Los Angeles, California.

[62] Yohan Jo, and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In Proceedings of the fourth ACM international conference on Web search and data mining, pages 815-824.

[63] Himabindu Lakkaraju, Richard Socher, and Chris Manning. 2014. Aspect Specific Sentiment Analysis using Hierarchical Deep Learning. NIPS Workshop on Deep Learning and Representation Learning, 2014.

[64] Li Zhuang, Feng Jing, Xiao yan Zhu, and Lei Zhang. 2006. Movie review mining and summarization. Proceedings of ACM International Conference on Information and Knowledge Management (CIKM-2006), 2006.

[65] Qui Guang, Liu Bing, Bu, J., Chen, C. 2011. Opinion Word Expansion and Target Extraction through Double Propagation. Computational Linguistics 37 (1).

[66] Fangtao Li, Chao Han, Minlie Huang, Xiaoyan Zhu, Ying-Ju Xia, Shu Zhang,and Hao Yu. (2010). Structure-Aware Review Mining and Summarization. International Conference on Computational Linguistics (COLING).

[67] Abdulaziz Alghunaim, MitraMohtarami, Scott Cyphers, and Jim Glass. 2015. A Vector Space Approach for Aspect Based Sentiment Analysis. Proceedings of the 1st Workshop on Vector SpaceModeling for Natural Language Processing, pages 116–122, Denver, Colorado, June. Association for Computational Linguistics.

[68] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In NIPS.

[69] Jeffery Pennington, Richard Socher, Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), pages 1532–1543, Doha, QA.

[70] Maria Pontiki and Haris Papageorgiou. 2015. Opinion Mining and Target Extraction in Greek Review Texts. Proceedings of the 12th International Conference on Greek Linguistics (ICGL12), Berlin, Germany.