Μακάριος Χρηστάκης

1046895

Α Έτος (2015-2016)

Συλλογή και ανάλυση δεδομένων στην πρόβλεψη κοινωνικών εξελίξεων.

ΠΕΡΙΛΗΨΗ

Η συλλογή και η ανάλυση δεδομένων μεγάλου όγκου είναι ένας σχετικά νέος κλάδος στην επιστήμη των υπολογιστών και της στατιστικής. Η διαδικασία επεξεργασίας των πληροφοριών αυτών πρέπει να είναι προσεκτική και αυστηρή, έτσι ώστε να φιλτραριστεί η άχρηστη πληροφορία και να εξαχθεί ένα ορθό συμπέρασμα που ανταποκρίνεται στην πραγματικότητα που ζούμε. Τα συμπεράσματα αυτά μπορούν να έχουν προληπτικό (εύρεση πιθανών κρουσμάτων ασθενειών) ή αναλυτικό χαρακτήρα (κατασκευή παγκόσμιων χαρτών για καταγραφή κινημάτων). Επίσης μπορούν να προβλέψουν στατιστικά το αποτέλεσμα κάποιων πολιτικών διαδικασιών (εκλογές). Πάραυτα, δεν υπάρχουν ακόμα αρκετές πρακτικές υλοποιήσεις ολοκληρωμένων συστημάτων συλλογής κι επεξεργασίας μεγάλου όγκου δεδομένων (Big Data) με κάποιο συγκεκριμένο στόχο.

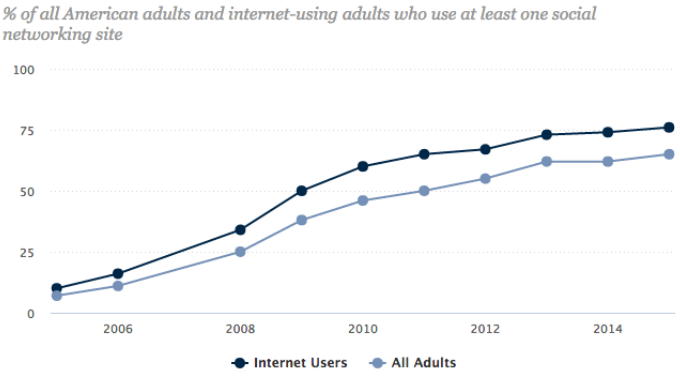
I. ΕΙΣΑΓΩΓΗ

Η εξόρυξη δεδομένων (Data Mining) είναι η διαδικασία εύρεσης μιας ενδιαφέρουσας, μη αυτονόητης και συχνά χρήσιμης πληροφορίας από μεγάλες βάσεις δεδομένων, χρησιμοποιώντας αλγόριθμους κατηγοριοποίησης και ομαδοποίησης δεδομένων και τις αρχές της στατιστικής. Στόχος της εξόρυξης δεδομένων είναι η πληροφορία που θα εξαχθεί και τα πρότυπα που θα προκύψουν να έχουν δομή κατανοητή προς τον άνθρωπο έτσι ώστε να τον βοηθήσουν να πάρει τις κατάλληλες αποφάσεις[1]. Η ανάλυση δεδομένων μπορεί να βοηθήσει τον άνθρωπο να εξάγει συμπεράσματα τόσο σε τοπικό, όσο και σε παγκόσμιο επίπεδο και να προβλέψει κοινωνικές εξελίξεις που ενδέχεται να συμβούν στους διάφορους πληθυσμούς. Προφανώς, το πόσο εύστοχα είναι αυτά τα συμπεράσματα εξαρτάται από τον αριθμό των δεδομένων που αναλύθηκαν, το πόσο αντιπροσωπευτικό είναι το δείγμα που χρησιμοποιήθηκε και τη διαδικασία/μέθοδο ανάλυσής τους. Στην εργασία αυτή παρουσιάζονται πορίσματα ερευνών πάνω σε διάφορα θέματα με κύριο χαρακτηριστικό την χρησιμοποίηση δεδομένων του διαδικτύου για την αναγνώριση και πρόβλεψη κοινωνικών μεταβολών.

II. ΑΝΑΠΤΥΞΗ ΘΕΜΑΤΟΣ

Το διαδίκτυο αποτελεί πλέον κομμάτι της καθημερινής ζωής του μέσου ανθρώπου και οι μηχανές αναζήτησης χρησιμοποιούνται κατά κόρον για την εύρεση πληροφοριών και την απάντηση ερωτημάτων που προκύπτουν κατά τη διάρκεια της ημέρας.

Επιπλέον, η ραγδαία άνοδος σε δημοτικότητα των κοινωνικών δικτύων τα τελευταία χρόνια (Σχήμα 1) ως μέσο επικοινωνίας και έκφρασης της κοινής γνώμης (πράγμα που αποτελεί επακόλουθο της γρήγορης εξέλιξης της τεχνολογίας, αφού είναι όλο και ευκολότερο πλέον να αποκτήσει κανείς «έξυπνες» φορητές συσκευές και πρόσβαση στο διαδίκτυο),



Σχήμα 1. Ποσοστό Αμερικανών ενηλίκων χρηστών του διαδικτύου που χρησιμοποιούν

τουλάχιστον ένα κοινωνικό δίκτυο από το 2005 μέχρι το 2015 (Οκτώβριος 2015) [2].

έχει ως αποτέλεσμα να υπάρχει ολοένα και μεγαλύτερος όγκος πληροφοριών αποθηκευμένων στις βάσεις δεδομένων των εκάστοτε κοινωνικών δικτύων.

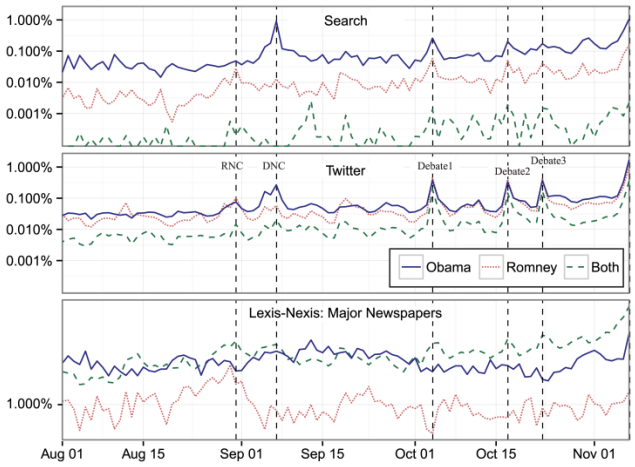
Συνεπώς, το διαδίκτυο είναι μια συνεχώς εμπλουτιζόμενη πηγή πληροφοριών για τα θέματα που απασχολούν την κοινωνία, τόσο σε τοπικό, όσο και σε παγκόσμιο επίπεδο. Παρακάτω θα γίνει αναφορά σε μελέτες, οι οποίες χρησιμοποίησαν στατιστικά μηχανών αναζήτησης (Google,Yahoo κλπ.), όπως και δεδομένα (δημοσιεύσεις, tweet κλπ.) από κοινωνικά δίκτυα.

A. Κοινωνικά Δίκτυα ως μέσο Δημοσκόπησης.

Τα κοινωνικά δίκτυα ενθαρρύνουν τους χρήστες τους να μοιράζονται την άποψη τους σε οποιοδήποτε θέμα τους απασχολεί. Συνεπώς η γνώμη ενός αρκετά μεγάλου μέρους του πληθυσμού για ένα συγκεκριμένο θέμα που απασχολεί την κοινωνία μπορεί να βρεθεί μέσω των δημοσιεύσεών στα κοινωνικά δίκτυα.

Για παράδειγμα, κατά τις Αμερικάνικες Προεδρικές Εκλογές του 2012 το διαδίκτυο και τα social networks ήταν γεμάτα από γνώμες για τους δυο υποψήφιους προέδρους, *Barrack Obama* και *Mitt Romney*. Σε μελέτη του 2016 [3], αναλύοντας τα ποσοστά των αναζητήσεων για συγκεκριμένες λέξεις-κλειδιά, την εμφάνισή των λέξεων αυτών σε δημοσιεύσεις του Twitter, όπως και την κάλυψη των θεμάτων αυτών στον γραπτό τύπο, φαίνεται πως η περίοδος απόσβεσης των συζητήσεων για το θέμα στον τύπο μετά τις κρίσιμες ημερομηνίες (ημερομηνίες Debate μεταξύ προέδρων) είναι μεγαλύτερη από αυτή στα social media. Συνεπώς οι πολίτες έχουν περιορισμένη διάρκεια προσοχής όσον αφορά τα πολιτικά ζητήματα.

Παρά όλα αυτά η μελέτη έδειξε πως ο νικητής των εκλογών *Barrack Obama* ήταν σταθερά υψηλότερα στις αναζητήσεις και τα tweet κατά τη διάρκεια της προεκλογικής περιόδου (Σχήμα 2).



Σχήμα 2. Ποσοστό του συνόλου των αναζητήσεων και συζήτησης στο Twitter και τον γραπτό τύπο για τους υποψήφιους προέδρους κατά την προεκλογική περίοδο (1η Αυγούστου με 6 Νοέμβρη 2012) [3].

Οπότε τα μέσα κοινωνικής δικτύωσης μπορούν να αποτελέσουν έναν σημαντικό δείκτη για το ενδιαφέρον του λαού και της κοινής γνώμης για κάποιο κοινωνικοπολιτικό θέμα. Έναν δείκτη διαφορετικό από αυτό των συμβατικών μέσων ενημέρωσης, όπως δημοσκοπήσεις στο τύπο και την τηλεόραση. Κατά μια έννοια, μπορεί να είναι ένας πιο «ζωντανός» δείκτης, αφού η συλλογή των δεδομένων γίνεται συνεχώς κι έτσι μπορεί να αλλάζει ανά πάσα στιγμή και πιο απότομα απ' ότι μια περιοδική δημοσκόπηση.

Αντιθέτως, είναι ένας αυστηρά υποκειμενικός δείκτης που βασίζεται στην ποσότητα των δημοσιεύσεων κι όχι το περιεχόμενό τους, πράγμα που τον κάνει εύκολα παρερμηνεύσιμο.

Γενικά, τα κοινωνικά δίκτυα μπορούν να χρησιμεύσουν ως τόποι για ζωντανές δημοσκοπήσεις μεγάλης κλίμακας, οι οποίες ανανεώνονται συνεχώς.

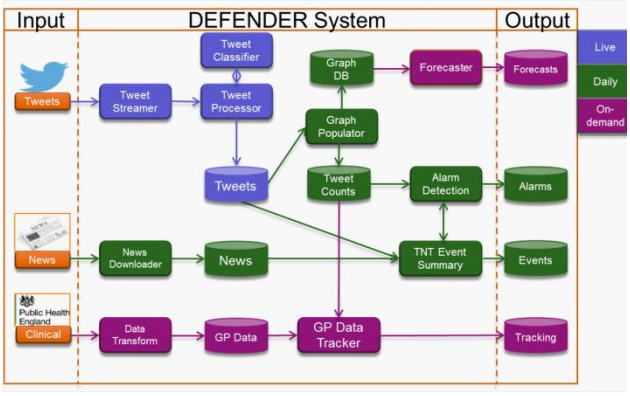
B. Πρόβλεψη πανδημιών.

Συλλέγοντας δεδομένα δεν μπορούμε να εξάγουμε μόνο συμπεράσματα για τις ιδέες των πολιτών, αλλά και για την κατάσταση της υγείας τους.

Με την σωστή επεξεργασία δεδομένων είναι εφικτό να έχουμε μια εικόνα για τις ασθένειες που απατώνται συχνά στις κοινωνίες, όπως και

να προβλέψουμε τυχόν πανδημίες. Έτσι, τα νοσοκομεία και κατ'επέκταση η ιατρική κοινότητα, έχοντας στη διάθεσή τους τις πληροφορίες αυτές, θα είναι σε θέση να προετοιμαστούν κατάλληλα για την αντιμετώπιση των ασθενειών.

Όλη η προεπεξεργασία των δεδομένων μέχρι να φτάσει σε ευκατανόητη και ερμηνεύσιμη μορφή, καθώς και το φιλτράρισμα του «θορύβου» (των άσχετων πληροφοριών που μαζεύονται δηλαδή), απαιτούν ένα ολοκληρωμένο σύστημα αλγορίθμων. Ήδη έχουν υπάρξει προτάσεις για τέτεια συστήματα, όπως το DEFENDER(Σχήμα 3), το οποίο επεξεργάζεται δεδομένα από εφημερίδες, κλινικές και Twitter για να προβλέψει, να ανιχνεύσει και να ειδοποιήσει για τυχόν επιδημίες ή εμφάνιση συμπτωμάτων σε μέρος του πληθυσμου.



Σχήμα 3. Η αρχιτεκτονική του συστήματος DEFENDER, που αποσκοπεί στην πρόληψη επιδημιών [4].

Τα συστήματα αυτά όμως δεν μπορούν ακόμα να θεωρηθούν έμπιστα, διότι παρ'όλο που ο όγκος των πληροφοριών που μαζεύεται από τα διάφορα δίκτυα είναι μεγάλος, δεν είναι αυτόματα και χρήσιμος, αλλά συχνά υποδεκαπλασιάζεται αφού περάσει από τα διάφορα φίλτρα κατηγοριοποίησης.

Η ανάλυση πληροφοριών για πρόβλεψη ξεσπασμάτων ασθενειών είναι ακόμα σε πρώιμο στάδιο, αλλά έχει προοπτικές να γίνει χρήσιμο εργαλείο πρόληψης στο μέλλον.

C. Άλλες εφαρμογές.

Η εκμετάλλευση μεγάλου μεγέθους πληροφοριών δεν χρειάζεται να έχει πάντα ως στόχο την πρόβλεψη κάποιας αλλαγής, αλλά μπορεί απλά να μας δώσει μια αίσθηση του τι συμβαίνει γύρω μας.

Παραδείγματος χάριν, μια μελέτη του 2016 [5] συγκέντρωσε δεδομένα από γεωγραφικά χαρακτηρισμένες (Geo-Tagged) εικόνες του διαδικτύου με τη λέξη κλειδί "protest" (διαδήλωση), οι οποίες δημοσιεύτηκαν καθ' όλη τη διάρκεια του 2013 και απεικόνισε στον παγκόσμιο χάρτη τις χώρες που είχαν τις περισσότερες διαδηλώσεις.

Συνεπώς, τα Big Data μπορούν να μας χρησιμεύσουν στον εντοπισμό προτύπων συμπεριφοράς (behavioral patterns) των κοινωνικών ομάδων και να μας δείξουν το πώς οι ομάδες αυτές εξελίσσονται και αλλάζουν στον άξονα του χρόνου.

III. ΣΥΝΟΨΗ

Στην εποχή που το διαδίκτυο και τα social media είναι ευρέως διαδεδομένα και χρησιμοποιούνται καθημερινά, η άντληση μεγάλου όγκου πληροφοριών δεν είναι δύσκολη. Οι πληροφορίες αυτές, έπειτα από σωστή επεξεργασία μπορούν να βοηθήσουν το ανθρώπινο είδος τόσο να κατανοήσει τις αλλαγές που συμβαίνουν γύρω του (ιδεολογικές, κοινωνικοπολιτικές), όσο και να προβλέψει μη αναμενόμενες μελλοντικές δυσκολίες (πχ. επιδημίες σε μεγάλα κομμάτια πληθυσμού), με αποτέλεσμα να είναι σε θέση να προετοιμαστεί έγκαιρα για να τις αντιμετωπίσει.

Παρ' όλα αυτά δεν πρέπει να αμελήσουμε το γεγονός πως οι πληροφορίες αυτές μπορούν εύκολα να παρερμηνευθούν με άμεσο επακόλουθο την εξαγωγή λανθασμένων συμπερασμάτων και την εσφαλμένη πράξη.

Σε γενικές γραμμές όμως, κι έπειτα από προσεκτική επεξεργασία και κατηγοριοποίηση, οι πληροφορίες μπορούν να μας οδηγήσουν σε μια καλύτερη κατανόηση του κοινωνικού μας περίγυρου, αλλά και γενικότερα του κόσμου.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1]. https://en.wikipedia.org/wiki/Data_mining

[2].http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/

[3]. Diaz F, Gamon M, Hofman JM, Kıcıman E, Rothschild D (2016) Online and Social Media Data As an Imperfect Continuous Panel Survey. PLoS ONE 11(1): e0145406. doi:10.1371/journal.pone.0145406

[4]. Thapen N, Simmie D, Hankin C, Gillard J (2016) DEFENDER: Detecting and Forecasting Epidemics Using Novel Data-Analytics for Enhanced Response. PLoS ONE 11(5): e0155417. doi:10.1371/journal.pone.0155417

[5]. Alanyali M, Preis T, Moat HS (2016) Tracking Protests Using Geotagged Flickr Photographs. PLoS ONE 11(3): e0150466. doi:10.1371/journal.pone.0150466