ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΠΑΓΚΟΣΜΙΟΥ ΙΣΤΟΥ

Α.Π.Θ. Τμήμα Πληροφορικής
Χειμερινό εξάμηνο 2016-2017

Θέμα εργασίας: Ανακάλυψη κοινοτήτων χρηστών στο Twitter

Αναπόσπαστο κομμάτι της καθημερινότητας των μελών της κοινωνίας του σήμερα αποτελεί η ενασχόλησή τους με τα κοινωνικά δίκτυα, όπως το Twitter. Σύμφωνα με επίσημα στατιστικά στοιχεία του 2ου τετράμηνου του 2016 συνολικά υπάρχουν 313M ενεργοί χρήστες. Οι χρήστες του Twitter συνήθως δημιουργούν σύντομα μηνύματα, γνωστά ως tweets, τα οποία μπορούν να περιλαμβάνουν κείμενο, υπερσυνδέσμους, ετικέτες επισήμανσης (hashtags), αναφορές σε άλλους χρήστες (mentions), εικόνες, καθώς και τη γεωγραφική θέση του χρήστη.

Βάσει της δραστηριότητας των χρηστών στο Twitter δημιουργούνται σχέσεις μεταξύ αυτών, είτε ρητές, είτε άρρητες. **Ρητές** είναι οι σχέσεις οι οποίες είτε δηλώνονται από τους ίδιους τους χρήστες, π.χ. ο χρήστης Α είναι ακόλουθος του χρήστη Β, είτε προκύπτουν από τις έμμεσες αλληλεπιδράσεις τους, π.χ. ο χρήστης Α αναφέρει στο post του (mention) το χρήστη Β. **Άρρητες** είναι οι σχέσεις που δε δηλώνονται άμεσα, αλλά μπορούν να εξαχθούν βάσει ομοιοτήτων που παρατηρούνται σε ορισμένες αλληλεπιδράσεις των χρηστών με κοινές οντότητες. Για παράδειγμα η χρήση κοινών hashtags από δύο χρήστες υποδηλώνει μία άρρητη σχέση μεταξύ τους η οποία αφορά πιθανόν σε κοινά ενδιαφέροντα. Τόσο οι ρητές όσο και οι άρρητες σχέσεις μεταξύ των χρηστών μπορούν να αξιοποιηθούν για την ομαδοποίηση αυτών σε κοινότητες τα μέλη των οποίων έχουν κοινά χαρακτηριστικά δραστηριότητας και θεματικά ενδιαφέροντα [1, 2].

Για τον εντοπισμό κοινοτήτων αρχικά απαιτείται ο προσδιορισμός ενός *μέτρου ομοιότητας* για οποιοδήποτε ζεύγος χρηστών το οποίο θα βασίζεται σε επιλεγμένα χαρακτηριστικά δραστηριότητας αυτών. Στη συνέχεια, μία προσέγγιση περιλαμβάνει τον υπολογισμό του *πίνακα ομοιότητας* (similarity matrix) για το σύνολο των χρηστών βάσει του επιλεγμένου μέτρου, και τη δημιουργία *γράφου* με τους χρήστες ως κόμβους και με το βαθμό ομοιότητας μεταξύ δύο χρηστών να εκφράζει το βάρος της αντίστοιχης ακμής. Όπως είναι προφανές στον παραχθέντα γράφο υπάρχουν ακμές μόνο μεταξύ των χρηστών που έχουν μηδενική ομοιότητα. Τέλος, πάνω στο γράφο μπορεί να γίνει εφαρμογή κάποιου αλγορίθμου ανεύρεσης κοινοτήτων (community detection) οι οποίοι επιτρέπουν το διαχωρισμό των κόμβων σε ομάδες βάσει των δομικών χαρακτηριστικών του γράφου και κάποιων μετρικών ποιότητας [3]. Στα πλαίσια των αλγορίθμων αυτών μία κοινότητα ορίζεται ως: *ένα σύνολο κόμβων οι οποίοι είναι πιο πυκνά συνδεδεμένοι μεταξύ τους σε σχέση με το υπόλοιπο δίκτυο*.

Σκοπός της εργασίας είναι η ανεύρεση κοινοτήτων χρηστών λαμβάνοντας υπόψη τις έμμεσες σχέσεις μεταξύ αυτών. Στη συνέχεια περιγράφονται αναλυτικά τα βήματα που θα πρέπει να ακολουθηθούν.

Αναλυτική περιγραφή

1ο μέρος: Συλλογή δεδομένων (έως 11/11/2016)

Θα γίνει χρήση του Twitter Streaming API[1], το οποίο επιτρέπει τη συνεχή παροχή νέων tweets βάσει κριτηρίων, για τη συλλογή δεδομένων που σχετίζονται με ένα θέμα της επιλογής σας. Για τη συγκέντρωση μεγάλου όγκου δεδομένων μία προσέγγιση αποτελεί η επιλογή θέματος από τη λίστα των δημοφιλών θεμάτων έτσι όπως αυτά χαρακτηρίζονται από το Twitter. Το Twitter υπολογίζει και ανανεώνει ανά 5 λεπτά λίστες με τα top-10 δημοφιλή θέματα (λέξεις-φράσεις κλειδιά) βάσει της συχνότητας αναφοράς σε αυτά σε διαφορετικές γεωγραφικές περιοχές ή παγκοσμίως.

2ο μέρος: Αποθήκευση tweets σε βάση δεδομένων (έως 11/11/2016)

Για την καλύτερη διαχείριση των δεδομένων που θα συλλεχθούν απαραίτητη είναι η χρήση βάσης δεδομένων η οποία θα επιτρέπει τη γρήγορη και εύκολη ανάκτησή τους. Στην πλαίσια της εργασίας θα χρησιμοποιηθεί η βάση δεδομένων MongoDB[2] (open source, NoSQL, document-oriented βάση δεδομένων). Η διατήρηση των tweets θα πρέπει να είναι στη μορφή που επιστρέφονται από το Twitter (JSON αναπαράσταση).

3ο μέρος: Ανάλυση κειμένων tweets (έως 05/12/2016)

Το 3ο μέρος της εργασίας περιλαμβάνει το διαχωρισμό του κειμένου σε λεξιλογικές μονάδες (tokens) και την αναγνώριση τυχόν hashtags, URLs, αναφορές σε άλλους χρήστες (mentions) ή σε άλλα tweets (retweets) που περιλαμβάνονται σε αυτό. Για το σκοπό αυτό θα πρέπει να ληφθεί υπόψιν ότι τα hashtags στο Twitter είναι όροι οι οποίοι ξεκινούν με τον χαρακτήρα #, τα mentions σε χρήστες υποδηλώνονται συμπεριλαμβάνοντας το username του χρήστη μετά από το χαρακτήρα @, και ότι τα retweets δηλώνονται με συγκεκριμένη δομή μέσα στην JSON αναπαράσταση του tweet.

Προσοχή θέλει η αναγνώριση των υπερσυνδέσμων που εμπεριέχονται στα tweets μιας και πολύ συχνά χρησιμοποιούνται συντμημένα URLs είτε μέσω της αντίστοιχης υπηρεσίας του Twitter είτε μέσω τρίτων υπηρεσιών, όπως είναι η υπηρεσία bit.ly. Στις περιπτώσεις εμφάνισης συντμημένων URLs θα πρέπει πρώτα να γίνει αναγωγή στην πλήρη τους μορφή (όπου αυτό καθίσταται δυνατό).

Μοντελοποίηση χαρακτηριστικών δραστηριότητας χρήστη. Στο βήμα αυτό θα πρέπει να δημιουργήσετε κατάλληλη δομή / βάση αποθήκευσης δεδομένων για τις προαναφερθείσες οντότητες, δηλαδή: hashtags, URLs (διατηρώντας την αντιστοίχιση της πλήρους και συντμηνένης μορφής), χρήστες και tweets.

Επιπλέον, θα πρέπει για κάθε χρήστη του συνόλου δεδομένων να αποθηκεύονται οι άμεσες συνδέσεις του με τις άλλες οντότητες (χρήστες, tweets, hashtags και URLs), αλλά και ο χρόνος χρήσης/αναφοράς (timestamp) της κάθε οντότητας. Οι πιθανές υφιστάμενες συνδέσεις για ένα χρήστη είναι οι εξής:

user, timestamp, hashtag
user, timestamp, URL
user, timestamp, mentioned_user
user, timestamp, retweeted_tweet

Για την αποθήκευση των δεδομένων αυτών θα πρέπει *να επιλέξετε την τεχνολογία / δομή που θεωρείτε καταλληλότερη δικαιολογώντας παράλληλα την επιλογή σας αυτή*. Πιθανές επιλογές είναι οι εξής: (i) κάποια document-oriented βάση δεδομένων, π.χ. MongoDB, CouchDB[3], (ii) κάποια βάση δεδομένων προσανατολισμένη σε γράφους, π.χ. Neo4j[4], ή (iii) κάποια σχεσιακή βάση δεδομένων, π.χ. MySQL[5].

4ο μέρος: Επιλογή μέτρων ομοιότητας και υπολογισμός της τιμής τους για κάθε ζεύγος χρηστών (έως 19/12/2016)

Στο σημείο αυτό θα πρέπει να επιλέξετε μέτρα ομοιότητας χρηστών τα οποία να βασίζονται στα κοινά χαρακτηριστικά δραστηριότητάς τους (π.χ. Cosine similarity, Jaccard similarity, απλά frequency-based,

κ.α.). Η επιλογή του μέτρου ομοιότητας θα πρέπει να δικαιολογηθεί με σχετική αναφορά στην αρθρογραφία.

Στη συνέχεια, θα πρέπει να γίνει υπολογισμός της τιμής του μέτρου για καθένα από τα 4 χαρακτηριστικά που αναφέρθηκαν παραπάνω, δηλαδή αναφορά σε κοινούς χρήστες, αναφορά σε κοινά tweets, χρήση κοινών hashtags, και χρήση κοινών URLs, και να δημιουργήσετε τον πίνακα ομοιότητας. Στο τέλος της διαδικασίας αυτής θα πρέπει να έχετε 5 πίνακες ομοιότητας, έναν για κάθε χαρακτηριστικό ξεχωριστά, και έναν επιπλέον για το συνδυασμό των μέτρων ομοιότητας σε ένα συνολικό μέτρο. Στην περίπτωση του συνολικού μέτρου θα πρέπει να προσέξετε μήπως χρειάζεται να γίνει κάποια **κανονικοποίηση** των τιμών των επιμέρους μέτρων πριν τον συνδυασμό τους.

Προσοχή. Είναι πολύ πιθανό να υπάρχουν στο σύνολο των δεδομένων οντότητες οι οποίες χρησιμοποιούνται από πολύ μεγάλο ποσοστό των χρηστών (π.χ. κάποιο γενικό hashtag). Οι οντότητες αυτές δε βοηθούν στη διάκριση των επιμέρους ενδιαφερόντων των χρηστών και συνεπώς καλό είναι να αφαιρούνται. Έτσι, αρχικά θα πρέπει να υπολογίσετε την εμπειρική κατανομή της χρήσης των οντοτήτων στο σύνολο των χρηστών και να αποφασίσετε ποιες θα πρέπει να αφαιρεθούν λόγω της χρήσης τους από πολύ μεγάλο αριθμό χρηστών (π.χ. μέσω της επιλογής κάποιου ποσοστού, της χρήσης κάποιας τεχνικής για την εύρεση outliers όπως είναι τα boxplots, κ.λ.π.).

5ο μέρος: Δημιουργία και απεικόνιση γράφου χρηστών (έως 26/12/2016)

Χρησιμοποιώντας τον κάθε πίνακα ομοιότητας θα πρέπει να σχηματίσετε τον αντίστοιχο γράφο χρηστών. Στον κάθε γράφο, μία ακμή ανάμεσα σε δύο χρήστες θα έχει ως βάρος την τιμή του αντίστοιχου μέτρου ομοιότητας που έχει υπολογιστεί γι' αυτούς. Καθώς η σύνδεση χρηστών που έχουν πολύ μικρή ομοιότητα με ακμή μπορεί να αυξήσει σημαντικά την πυκνότητα του γράφου, και επιπλέον δεν υποδηλώνει την ύπαρξη σημαντικής συνάφειας στα ενδιαφέροντα / απόψεις των χρηστών, καλείστε να φιλτράρετε τέτοια ζεύγη χρηστών και να μη τα συνδέσετε στο γράφο. Για το σκοπό αυτό μπορείτε να ακολουθήσετε κάποια μέθοδο ανάλογη με αυτήν που περιγράφηκε στο 4ο μέρος της εργασίας για τις πολύ συχνές οντότητες.

Για την απεικόνιση των γράφων που θα δημιουργήσετε προτείνεται το εργαλείο Gephi[6]. Για την εισαγωγή των γράφων σε αυτό θα πρέπει πρώτα να τους αναπαραστήσετε σε κάποιο από τα formats που υποστηρίζει. Επιπλέον, προαιρετικά οι γράφοι μπορούν να αποθηκευτούν και στη βάση δεδομένων που αναπτύξατε στο 3ο μέρος της εργασίας. Αφού εισάγετε τους γράφους στο Gephi, δοκιμάστε τα διάφορα layouts που διαθέτει για την οπτικοποίησή τους και παρατηρήστε τυχόν διαφορές / ομοιότητες.

6ο μέρος: Υπολογισμός στατιστικών μεγεθών και ανεύρεση κοινοτήτων (έως 02/01/2017)

Το Gephi δίνει τη δυνατότητα υπολογισμού διάφορων μετρικών (Measures) που σχετίζονται με τη δομή και τη συνεκτικότητα του γράφου. Στο βήμα αυτό θα πρέπει με τη χρήση αυτής της επιλογής να υπολογίσετε μέτρα όπως: (i) avg clustering coefficient, (ii) diameter, (iii) graph density και να σχολιάσετε συγκριτικά τα αποτελέσματα για τους γράφους. Στη συνέχεια, θα εφαρμόσετε σε κάθε γράφο τον αλγόριθμο ανεύρεσης κοινοτήτων Louvain [4] που προσφέρεται από το Gephi (εμφανίζεται ως modularity). Κατά την εκτέλεση του αλγορίθμου αυτού θα πρέπει να λάβετε υπόψιν τα βάρη των ακμών. Έχοντας υπολογίσει τις κοινότητες, επιλέξτε το χρωματισμό των κόμβων ανάλογα με την κοινότητα που ανήκει ο εκάστοτε χρήστης, καθώς και το κατάλληλο layout και παρατηρήστε τις δομές που αναδύθηκαν από τους γράφους και την τιμή του μέτρου ποιότητας *modularity*.

7ο μέρος: Σύγκριση των δομών κοινοτήτων (έως 16/01/2017)

Στο τελευταίο μέρος της εργασίας θα πρέπει να εξάγετε από το Gephi το αρχείο που περιγράφει την ανάθεση των κόμβων σε κοινότητες για κάθε γράφο. Στόχος είναι η παρατήρηση της ομοιότητας / διαφοράς που παρουσιάζουν οι ομαδοποιήσεις που έγιναν με τα διαφορετικά μέτρα ομοιότητας. Για το σκοπό αυτό θα υπολογίσετε ανά ζεύγος μέτρων το δείκτη NMI (Normalized Mutual Information) [5] για τις κοινότητες που ανακαλύφθηκαν και θα σχολιάσετε τα αποτελέσματα.

Η εργασία θα εκπονηθεί από ομάδες μέχρι 4 άτομα.

Παράδοση εργασίας

Κάθε ομάδα θα πρέπει να παραδώσει ένα συμπιεσμένο φάκελο με όνομα τα ΑΕΜ των φοιτητών της ομάδας (π.χ. 1234_2341_3412.zip). Ο φάκελος θα περιλαμβάνει τα εξής:

1. Τον **πηγαίο κώδικα Java** που αναπτύχθηκε για τη συλλογή, την προεπεξεργασία, τη μοντελοποίηση και αποθήκευση των δεδομένων, καθώς και για τον υπολογισμό των μέτρων ομοιότητας και τη σύγκριση των δομών κοινοτήτων.
2. Μία **αναφορά** (~5-10 σελίδες) η οποία θα περιλαμβάνει: (i) την περιγραφή του μοντέλου που χρησιμοποιήθηκε για την αναπαράσταση των δεδομένων, (ii) την περιγραφή της υλοποίησης των μεθόδων επεξεργασίας και υπολογισμού, (iii) το σχολιασμό των αποτελεσμάτων και την τεκμηρίωση των επιλογών που ζητούνται στα διάφορα βήματα, και (iv) τα ζητούμενα διαγράμματα και απεικονίσεις.
3. **Αρχεία εισόδου δεδομένων** στο Gephi και τα αντίστοιχα **Gephi projects**.

Χρήσιμοι σύνδεσμοι

1. Twitter Streaming API: https://dev.twitter.com/streaming/overview
2. MongoDB: https://www.mongodb.com/
3. CouchDB: http://couchdb.apache.org/
4. Neo4j: https://neo4j.com/
5. MySQL: https://www.mysql.com/
6. Gephi: https://gephi.org/

Σχετική αρθρογραφία

1. Davide Frey, Arnaud Jégou, and Anne-Marie Kermarrec. 2011. Social market: combining explicit and implicit social networks. In Proceedings of the 13th international conference on Stabilization, safety, and security of distributed systems (SSS'11), Xavier Défago, Franck Petit, and Vincent Villain (Eds.). Springer-Verlag, Berlin, Heidelberg, 193-207.
2. Folke Mitzlaff, Martin Atzmueller,Dominik Benz, Andreas Hotho, Gerd Stumme: Community Assessment Using Evidence Networks. MSM/MUSE 2010: 79-98.
3. Santo Fortunato. Community detection in graphs. Eprint arXiv: 0906.0612. Physics Reports 486, 75-174 (2010).
4. Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P1000.
5. Danon, L.; Diaz-Guilera, A.; Duch, J.; and Arenas, A. 2005. Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment P09008(0505245).

Παράδοση εργασίας: 16 Ιανουαρίου 2017