ΣΥΣΤΗΜΑ ΑΥΤΟΜΑΤΗΣ ΕΞΑΓΩΓΗΣ ΑΛΛΗΛΕΠΙΔΡΑΣΕΩΝ MICRORNA-ΓΟΝΙΔΙΩΝ ΑΠΟ ΣΥΝΟΔΕΥΤΙΚΟ ΥΛΙΚΟ ΕΠΙΣΤΗΜΟΝΙΚΩΝ ΔΗΜΟΣΙΕΥΣΕΩΝ

ΠΛΗΡΟΦΟΡΙΕΣ: Θανάσης Βεργούλης, vergoulis@imis.athena-innovation.gr, Ηλίας Κανέλλος, kanellos@dblab.ece.ntua.gr, Θοδωρής Δαλαμάγκας, dalamag@imis.athena-innovation.gr

ΠΕΡΙΛΗΨΗ: Η μελέτη των *βιομοριών* που συμμετέχουν στους μηχανισμούς της ζωής (πχ DNA, πρωτεΐνες, μόρια microRNA κτλ) είναι απαραίτητη για να μπορέσουν οι ερευνητές να κατανοήσουν και να θεραπεύσουν γενετικές ασθένειες που απασχολούν την Ιατρική και τη Βιολογία τους τελευταίους αιώνες. Οι πληροφορίες που σχετίζονται με αυτά τα βιομόρια αποκαλύπτονται μέσω βιολογικών πειραμάτων ή υπολογιστικών προβλέψεων και τα αποτελέσματα δημοσιεύονται σε επιστημονικά περιοδικά και συνέδρια. Για να καταστεί η πληροφορία που καταγράφεται σε αυτές τις δημοσιεύσεις εύκολα προσβάσιμη στους βιοεπιστήμονες ανατίθεται σε επιμελητές η ανάγνωση της βιβλιογραφίας και η καταγραφή της γνώσης που εντοπίζεται με συστηματικό τρόπο. Επειδή όμως η συγκεκριμένη εργασία έχει αποδειχθεί ιδιαίτερα χρονοβόρα και επίπονη, κρίθηκε απαραίτητη η υλοποίηση εργαλείων που (α) βοηθούν τον επιμελητή να εντοπίσει γρήγορα τις δημοσιεύσεις που τον αφορούν και (β) σκιαγραφούν το είδος της πληροφορίας που καταγράφεται σε καθεμία από τις δημοσιεύσεις.
Σε προηγούμενες εργασίες έχουμε υλοποιήσει μοντέλα που μπορούν να χρησιμοποιηθούν για την αυτόματη εξαγωγή αλληλεπιδράσεων μορίων miRNA με γονίδια χρησιμοποιώντας το κείμενο επιστημονικών δημοσιεύσεων. Όμως, η συγκεκριμένη μεθοδολογία αποτυγχάνει να εντοπίσει ένα μεγάλο μέρος των αλληλεπιδράσεων οι οποίες περιέχονται σε συνοδευτικό υλικό των εργασιών όπως είναι πίνακες και εικόνες. Σκοπός της παρούσας εργασίας είναι η υλοποίηση ενός συστήματος που θα επιδιώκει την εξαγωγή σημαντικού μέρους των προαναφερθεισών αλληλεπιδράσεων.

ΑΤΟΜΑ: 1

ΠΛΑΤΦΟΡΜΑ ΕΡΓΑΣΙΑΣ: Python, Java, MySQL.

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ: Το σύνολο της γενετικής πληροφορίας ενός οργανισμού κωδικοποιείται σε ακολουθίες DNA, που ονομάζονται *γονίδια*. Το κύτταρο «διαβάζει» τη γενετική πληροφορία που κωδικοποιούν τα γονίδια και, με βάση αυτή, παράγει *πρωτεΐνες*, θέτοντας έτσι σε εφαρμογή τους μηχανισμούς της ζωής. Δυσλειτουργίες κατά την παραγωγή πρωτεϊνών μπορούν να δημιουργήσουν προβλήματα στους μηχανισμούς αυτούς. Τέτοιες δυσλειτουργίες αποτελούν την αιτία πολλών γενετικών ασθενειών. Τα μόρια microRNA, τα οποία λειτουργούν ως ρυθμιστές της παραγωγής πρωτεϊνών, υπόσχονται την παροχή ενός τρόπου αντιμετώπισης τέτοιων ασθενειών.
Λόγω όσων αναφέρθηκαν προηγουμένως, οι πληροφορίες που σχετίζονται με γονίδια και μόρια microRNA είναι ιδιαίτερα χρήσιμες στους ερευνητές. Τέτοιες πληροφορίες προκύπτουν από την εκτέλεση βιολογικών πειραμάτων ή από προγνώσεις υπολογιστικών προσομοιώσεων και καταγράφονται σε επιστημονικές εργασίες που δημοσιεύονται σε περιοδικά. Θεωρητικά η δημοσίευση των εργασιών παρέχει πρόσβαση στην επιστημονική γνώση για τους ερευνητές όλου του κόσμου, όμως, στην πράξη, το πλήθος των δημοσιεύσεων είναι τόσο μεγάλο που είναι δύσκολο ένας ερευνητής να είναι ενήμερος για όλη τη σχετική βιβλιογραφία. Για να βελτιωθεί η κατάσταση, ανατίθεται σε επιμελητές η ανάγνωση της βιβλιογραφίας και η καταγραφή της γνώσης που εντοπίζεται με τρόπο συστηματικό. Όμως και πάλι η εργασία των επιμελητών είναι χρονοβόρα και κοπιαστική. Για να διευκολυνθεί η εργασία αυτή είναι απαραίτητη η χρήση εργαλείων που

βοηθούν τους επιμελητές να εντοπίσουν γρήγορα τις δημοσιεύσεις που τους αφορούν και να εξάγουν με εύκολο τρόπο την πληροφορία που περιέχεται σε αυτές τις δημοσιεύσεις.
Σε προηγούμενη εργασία έχουμε αναπτύξει ένα μοντέλο για την αυτόματη εξαγωγή γνώσης από κείμενα επιστημονικών δημοσιεύσεων. Η εργασία αυτή στηρίζεται σε τεχνικές επεξεργασίας φυσικής γλώσσας, επιτυγχάνει τον εντοπισμό μεγάλου μέρους των αλληλεπιδράσεων που αναφέρονται σε προτάσεις του κειμένου, όμως αποτυγχάνει να εντοπίσει όσες αλληλεπιδράσεις περιέχονται μέσα σε πίνακες, εικόνες ή άλλο συνοδευτικό υλικό των δημοσιεύσεων. Σκοπός είναι να αναπτυχθεί ένα σύστημα που θα μπορεί να εντοπίσει όσο το δυνατόν περισσότερες από τις αλληλεπιδράσεις που περιέχονται σε συνοδευτικό υλικό δημοσιεύσεων.
Η παρούσα εργασία περιλαμβάνει:

- Διερεύνηση των δυνατοτήτων που υπάρχουν για αυτόματη εξαγωγή γνώσης από εικόνες και πίνακες που περιέχονται σε επιστημονικές δημοσιεύσεις για βιομόρια microRNA (πχ χρήση αλγορίθμων OCR ή αλγορίθμων για parsing πινάκων κτλ),
- Υλοποίηση αλγορίθμων που χρησιμοποιούν τα αποτελέσματα της προηγούμενης διερεύνησης προκειμένου να εξάγουν γνώση σχετική με τα βιομόρια microRNA

ΣΧΕΤΙΚΟ ΥΛΙΚΟ:

- Σύστημα αυτόματης εξαγωγής αλληλεπιδράσεων μεταξύ μορίων microRNA και γονιδίων από επιστημονικές βάσεις δεδομένων (διπλωματική εργασία Ροδοθέας-Μυρσίνης Τσουπίδη): http://www.dbnet.ece.ntua.gr/pubs/uploads/DIPL-2014-6.pdf

ΕΞΑΓΩΝΤΑΣ ΧΑΡΤΕΣ ΑΠΟ ΕΝΑΝ ΩΚΕΑΝΟ ΔΕΔΟΜΕΝΩΝ

ΠΛΗΡΟΦΟΡΙΕΣ: Καραγιώργου Σοφία, 210 6875430, karagior@imis.athena-innovation.gr

ΠΕΡΙΛΗΨΗ: Οι τεχνολογίες εντοπισμού θέσης που συναντώνται σε εφαρμογές πραγματικού χρόνου πλοήγησης ή διαχείρισης στόλου, όχι μόνο επιτρέπουν την παροχή μιας σημαντικής και συστηματικής πηγής δεδομένων παρακολούθησης που δίνει τη δυνατότητα να αντλήσει κανείς πληροφορία σχετικά με μεταφορικά δίκτυα, αλλά εισάγουν επίσης ένα πλήθος ανοιχτών ερευνητικών προκλήσεων σε σχέση με την αξιοποίηση τέτοιων πλούσιων δεδομένων. Στη βιβλιογραφία έχουν προταθεί ορισμένες μέθοδοι, ως μέθοδοι αυτόματης παραγωγής χαρτών, αλλά δεν έχουν μελετηθεί επαρκώς τα προβλήματα που απορρέουν από μεγάλης κλίμακας δεδομένα και γενικευμένης γεωμετρίας μεταφορικά δίκτυα. Στόχος της διπλωματικής εργασίας είναι: (α) Η ανάπτυξη αλγορίθμου αυτόματης παραγωγής χάρτη και εξαγωγής γνώσης για θαλάσσια δίκτυα (β) Η αξιοποίηση αυτού του αλγορίθμου πάνω από ένα μεγάλο όγκο δεδομένων σε παράλληλο και κατανεμημένο περιβάλλον.

ΑΤΟΜΑ: 1-2

ΠΛΑΤΦΟΡΜΑ ΕΡΓΑΣΙΑΣ: Java, MapReduce, MATLAB.

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ: Στόχος της παρούσης εργασίας είναι η ανάπτυξη αλγορίθμων εξαγωγής γνώσης από τροχιές κινούμενων πλοίων με αποτελεσματικό και βέλτιστο τρόπο σε επίπεδο α) χρόνου εκτέλεσης β) ποιότητας αποτελεσμάτων, δηλαδή εάν όντως ανακαλύπτουν σημαντικά χωρο-χρονικά στοιχεία της πορείας των πλοίων.
Τα καθήκοντα της εργασίας είναι τα παρακάτω:

- Ορισμός των μετρήσιμων χαρακτηριστικών της απόδοσης των αλγόριθμων, δηλαδή τι είναι αυτό που χαρακτηρίζει αποτελεσματικά την απόδοση και την αποτελεσματικότητα ενός αλγόριθμου αυτόματης εξαγωγής χάρτη.
- Ανάλυση, σχεδιασμός και αξιοποίηση μεγάλου όγκου χωρο-χρονικών δεδομένων για την εξόρυξη γεωμετρικών και επιπρόσθετων εξαγώμενων χαρακτηριστικών από τροχιές πλοίων.
- Η δημιουργία ενός θαλάσσιου δικτύου που θα χρησιμοποιηθεί για την αξιολόγηση και θα στηρίζεται στο OpenStreetMap ή συμπληρωματικές πηγές.

ΣΧΕΤΙΚΟ ΥΛΙΚΟ:

- Map construction - http://www.mapconstruction.org/
- Trajectory data mining - http://research.microsoft.com/apps/pubs/?id=241453
- Computing with Spatial Trajectories - http://research.microsoft.com/apps/pubs/default.aspx?id=153826
- MapReduce - https://en.wikipedia.org/wiki/MapReduce

ΕΞΑΓΩΝΤΑΣ ΧΑΡΤΕΣ ΑΠΟ ΣΥΝΝΕΦΑ ΔΕΔΟΜΕΝΩΝ

ΠΛΗΡΟΦΟΡΙΕΣ: Καραγιώργου Σοφία, 210 6875430, karagior@imis.athena-innovation.gr

ΠΕΡΙΛΗΨΗ: Οι τεχνολογίες εντοπισμού θέσης που συναντώνται σε εφαρμογές πραγματικού χρόνου πλοήγησης ή διαχείρισης στόλου, όχι μόνο επιτρέπουν την παροχή μιας σημαντικής και συστηματικής πηγής δεδομένων παρακολούθησης που δίνει τη δυνατότητα να αντλήσει κανείς πληροφορία σχετικά με μεταφορικά δίκτυα, αλλά εισάγουν επίσης ένα πλήθος ανοιχτών ερευνητικών προκλήσεων σε σχέση με την αξιοποίηση τέτοιων πλούσιων δεδομένων. Στη βιβλιογραφία έχουν προταθεί ορισμένες μέθοδοι, ως μέθοδοι αυτόματης παραγωγής χαρτών, αλλά δεν έχουν μελετηθεί επαρκώς τα προβλήματα που απορρέουν από μεγάλης κλίμακας δεδομένα και γενικευμένης γεωμετρίας μεταφορικά δίκτυα. Στόχος της διπλωματικής εργασίας είναι: (α) Η ανάπτυξη αλγορίθμου αυτόματης παραγωγής χάρτη και εξαγωγής γνώσης για δίκτυα εναέριας κυκλοφορίας (β) Η αξιοποίηση αυτού του αλγορίθμου πάνω από ένα μεγάλο όγκο δεδομένων σε παράλληλο και κατανεμημένο περιβάλλον.

ΑΤΟΜΑ: 1-2

ΠΛΑΤΦΟΡΜΑ ΕΡΓΑΣΙΑΣ: Java, MapReduce, MATLAB.

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ: Στόχος της παρούσης εργασίας είναι η ανάπτυξη αλγορίθμων εξαγωγής γνώσης από τροχιές κινούμενων αεροπλάνων με αποτελεσματικό και βέλτιστο τρόπο σε επίπεδο α) χρόνου εκτέλεσης β) ποιότητας αποτελεσμάτων, δηλαδή εάν όντως ανακαλύπτουν σημαντικά χωρο-χρονικά στοιχεία της πορείας των αεροπλάνων.
Τα καθήκοντα της εργασίας είναι τα παρακάτω:

- Ορισμός των μετρήσιμων χαρακτηριστικών της απόδοσης των αλγόριθμων, δηλαδή τι είναι αυτό που χαρακτηρίζει αποτελεσματικά την απόδοση και την αποτελεσματικότητα ενός αλγόριθμου αυτόματης εξαγωγής χάρτη.
- Συλλογή, ανάλυση, σχεδιασμός και αξιοποίηση μεγάλου όγκου χωρο-χρονικών δεδομένων για την εξόρυξη γεωμετρικών και επιπρόσθετων εξαγώμενων χαρακτηριστικών από τροχιές αεροπλάνων.
- Η δημιουργία ενός εναέριου δικτύου που θα χρησιμοποιηθεί για την αξιολόγηση του αλγορίθμου.

ΣΧΕΤΙΚΟ ΥΛΙΚΟ:

- Map construction - http://www.mapconstruction.org/
- Trajectory data mining - http://research.microsoft.com/apps/pubs/?id=241453
- Computing with Spatial Trajectories - http://research.microsoft.com/apps/pubs/default.aspx?id=153826
  MapReduce - https://en.wikipedia.org/wiki/MapReduce

ΕΝΣΩΜΑΤΩΣΗ ΑΝΟΙΚΤΩΝ ΓΕΩΓΡΑΦΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΣΤΟΝ ΣΗΜΑΣΙΟΛΟΓΙΚΟ ΙΣΤΟ

ΠΛΗΡΟΦΟΡΙΕΣ: Κώστας Πατρούμπας, 210 772 1446, kpatro@dblab.ece.ntua.gr

ΠΕΡΙΛΗΨΗ: Στόχος της εργασίας είναι η επέκταση της υπάρχουσας εφαρμογής *TripleGeo* με πρόσθετες δυνατότητες, ώστε να ανταποκρίνεται στις απαιτήσεις μετατροπής μεγάλης κλίμακας και ποικιλίας ανοικτών γεωγραφικών δεδομένων για την ενσωμάτωσή τους στον Σημασιολογικό Ιστό.

ΑΤΟΜΑ: 1

ΠΛΑΤΦΟΡΜΑ ΕΡΓΑΣΙΑΣ: Java, RDF, REST API.

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ: Χάρη στο μοντέλο λειτουργίας του *Σημασιολογικού Ιστού* (Semantic Web), έχουν αναπτυχθεί τεχνικές και εργαλεία για την διασύνδεση ανοικτών δεδομένων που αφορούν μεν την ίδια οντότητα, αλλά τα οποία πιθανόν προέρχονται από πολλαπλές, ανεξάρτητες και μάλλον ετερογενείς πηγές. Είναι χαρακτηριστικό ότι αυτού του είδους οι πληροφορίες συνήθως εμπεριέχουν (ρητώς ή εμμέσως) μία *γεωγραφική* πτυχή. Λ.χ. όταν ένας χρήστης ψάχνει για κάποια ταινία στο Διαδίκτυο ή απ' το κινητό του, η εφαρμογή θα μπορούσε να επιστρέφει όχι μόνο φωτογραφίες, κριτικές ή σχόλια για την ταινία, αλλά επίσης τις ώρες προβολής της σε κοντινούς κινηματογράφους μαζί με την τοποθεσία τους πάνω σε χάρτη.

Για να διευκολυνθεί η ενσωμάτωση τέτοιων *διασυνδεδεμένων γεωγραφικών δεδομένων* (*linked geospatial data*) στον Σημασιολογικό Ιστό, το ΙΠΣΥ έχει ξεκινήσει την ανάπτυξη της πλατφόρμας ***TripleGeo*** σε *ανοικτό κώδικα*. Στην τρέχουσα έκδοσή της, η εφαρμογή επιτρέπει την άντληση γεωγραφικών και περιγραφικών στοιχείων από υπάρχουσες πηγές (λ.χ. αρχεία ή βάσεις δεδομένων) και την εξαγωγή τους σε διάφορες μορφές κατάλληλες για τήρηση σε *αποθετήρια* (RDF stores) στο Διαδίκτυο.

Στα πλαίσια της διπλωματικής, θα επιδιωχθεί ο εμπλουτισμός της TripleGeo με πρόσθετες λειτουργίες επεξεργασίας, καθώς και με δυνατότητες χειρισμού μεγάλου πλήθους δεδομένων. Ενδεικτικά προτείνονται:

- Ανάπτυξη web interface και επέκταση των υπαρχουσών επιλογών εισόδου/εξόδου, λ.χ. επιτρέποντας ανάκτηση γεωγραφικών δεδομένων από αρχεία KML, GML κ.ά., καθώς και απ' ευθείας εισαγωγής των αποτελεσμάτων σε συγκεκριμένο αποθετήριο (λ.χ. Virtuoso RDF store).
- Υλοποίηση ενός περιβάλλοντος RESTful API, ώστε να είναι δυνατή η κλήση της εφαρμογής για ανοικτά γεωγραφικά δεδομένα προσβάσιμα μόνο μέσω Διαδικτύου.
- Σχεδιασμός και υλοποίηση μηχανισμού παράλληλης επεξεργασίας, λ.χ. με επιμερισμό του όγκου των στοιχείων για εκτέλεση από ανεξάρτητες εικονικές μηχανές (Virtual Machines), προκειμένου να αντιμετωπίζονται κλιμακούμενοι όγκοι δεδομένων (λ.χ. όλο το οδικό δίκτυο της Ευρώπης).
- Εφαρμογή της πλατφόρμας για την μετατροπή γεωγραφικών δεδομένων από ανοικτές πηγές (λ.χ. OpenStreetMap) και δυνατότητα εκθέσεώς τους στο Διαδίκτυο μέσω SPARQL endpoints.
- Μετρήσεις επιδόσεων της επεξεργασίας για διάφορους όγκους γεωγραφικών δεδομένων.

ΣΧΕΤΙΚΟ ΥΛΙΚΟ:

- TripleGeo: https://web.imis.athena-innovation.gr/redmine/projects/geoknow_public/wiki/TripleGeo
- REST API tutorial: http://rest.elkstein.org/2008/02/what-is-rest.html
- OpenStreetMap (OSM): http://www.openstreetmap.org/
- SPARQL protocol: http://www.w3.org/TR/rdf-sparql-protocol/

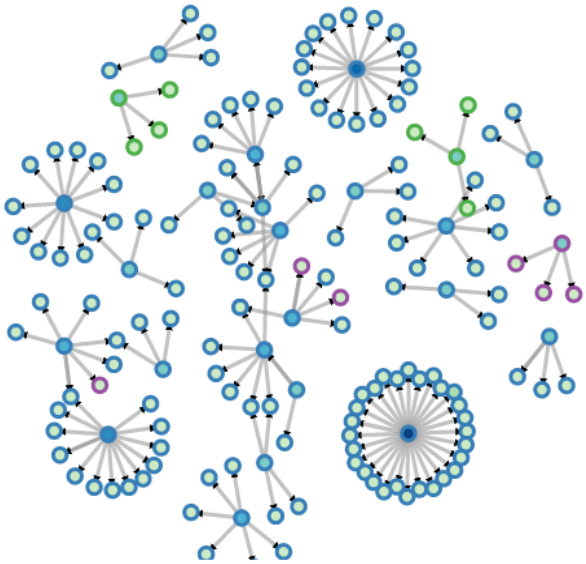
ΠΕΡΙΗΓΗΣΗ ΣΤΟ ΓΡΑΦΟ ΤΟΥ TWITTER

ΠΛΗΡΟΦΟΡΙΕΣ: Γιάννης Σταύρακας, yannis@imis.athena-innovation.gr
Δανάη Πλα Καρύδη, danae@imis.athena-innovation.gr
Αθανάσιος Γεντίμης, agent@imis.athena-innovation.gr

ΠΕΡΙΛΗΨΗ: Η περιήγηση στον γράφο των retweets, αποτελεί κομβική διαδικασία στον εντοπισμό της πορείας διάχυσης ενός ή περισσότερων tweets.
Στόχος της παρούσης εργασίας είναι: (α) Η ανάπτυξη μια ενιαίας πλατφόρμας που θα μπορεί να χρησιμοποιηθεί για την παρακολούθηση της διάδοσης tweets στον γράφο των χρηστών (β) Να χρησιμοποιηθεί αυτή η πλατφόρμα ως μέσο εξόρυξης χρήσιμης πληροφορίας από το Twitter συνδυάζοντας πληροφορίες από τα tweets για τον εμπλουτισμό των δυνατοτήτων περιήγησης

ΑΤΟΜΑ: 1-2

ΠΛΑΤΦΟΡΜΑ ΕΡΓΑΣΙΑΣ: Java, Neo4J, Gephi.



ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ: Οι χρήστες του Twitter συχνά επιλέγουν ορισμένα tweets άλλων χρηστών, για να τα διαδόσουν στους δικούς τους followers. Στόχος της παρούσης εργασίας είναι η ανάπτυξη κατάλληλων εργαλείων, τα οποία θα προσφέρουν στον χρήστη τη δυνατότητα να παρακολουθήσει για ένα retweet α) την πηγή του, β) την πορεία διάδοσής του και γ) τη συσχέτιση του με άλλα tweets μέσα από τα metadata των retweets ωστε η περιήγησή μας να εμπλουτίζεται από άποψη πληροφορίας. Σημαντικό κομμάτι της εργασίας αποτελεί η δημιουργία μιας εφαρμογής που θα αξιοποιεί ελληνικά tweets και η κατάλληλη προσαρμογή του προκειμένου να χρησιμοποιηθεί σε υπάρχοντα συστήματα του ΙΠΣΥ.
Συγκεκριμένα η εφαρμογή αρχικά θα εντοπίζει με βάση ένα tweet τον γράφο των retweets του. Ο γράφος θα περιλαμβάνει όλους τους χρήστες που αναπαρήγαγαν το tweet σε ένα συγκεκριμένο παράθυρο χρόνου (όσο το δυνατόν μεγαλύτερο). Ο χρήστης θα μπορεί να περιηγηθεί στους κόμβους βλέποντας επιπλέον πληροφορία για τα tweets του (τοποθεσία, τελευταία tweets, ημερομηνία). Τέλος αξιοποιώντας τα metadata (πχ URIs, εικόνες,βίντεο στα σχόλια) των retweets θα υπάρχει η δυνατότητα εύρεσης άλλων tweets που αναφέρονται στο ίδιο web περιεχόμενο. Αυτά τα tweets καθώς και ο γράφος των retweets τους θα μπορεί να εμπλουτίζει τον αρχικό γράφο.

ΣΧΕΤΙΚΟ ΥΛΙΚΟ:

- What is Twitter, a social network or a news media?- http://dl.acm.org/citation.cfm?id=1772751
- Generating graphs of retweets and @-messages on Twitter using R and Gephi - http://www.r-bloggers.com/generating-graphs-of-retweets-and-messages-on-twitter-using-r-and-gephi/