**Μάθημα:** [**Ευφυής Αλληλεπίδραση με Κοινωνικά Δίκτυα**](http://gunet2.cs.unipi.gr/courses/TMC132/)

**Ομάδα Εργασίας : Λιονής Χαράλαμπος ΜΠΣΠ19029**

**Βασίλης Ζωγράφος ΜΠΣΠ19013**

**Ημερομηνία: 16 - 07 - 2020**

1. Σκοπός εργασίας

Σκοπός της συγκεκριμένης εργασίας είναι η προγραμματιστική διαχείριση του χρονικά μεταβαλλόμενου δικτύου **Stack Overflow Temporal Network**. Πιο συγκεκριμένα μέσω των υπολογισμών μέτρων κεντρικότητας του παραπάνω δικτύου θα γίνει πρόβλεψη ακμών του γράφου.

1. Δεδομένα

Τα δεδομένα που χρησιμοποιήθηκαν βρέθηκαν στην διαδικτυακή τοποθεσία, https://snap.stanford.edu/data/sx-stackoverflow.html. Ο γράφος είναι αποθηκευμένος σε αρχείο κειμένου όπου αποτελείται από τρείς στήλες:

|  |  |
| --- | --- |
| source\_id | αναγνωριστικό του κόμβου προέλευσης της ακμής |
| target\_id | αναγνωριστικό του κόμβου κατάληξης της ακμής |
| timestamp | χρονική στιγμή δημιουργίας της ακμής |

1. Υλοποίηση – Υπολογισμοί

Για την εργασία χρησιμοποιήθηκε η γλώσσα προγραμματισμού python με networkx. Τα δεδομένα, αρχικά και παραγόμενα, βρίσκονται μέσα στον φάκελο data.

Αρχικά υλοποιήθηκε script (publicConstants.py) για τον ορισμό των σταθερών.

* Αριθμός τάξεων που θα δημιουργηθούν (number\_of\_classes = 10)
* Πλήθος στοιχείων σε κάθε κλάση (size\_of\_classes = 7000)

Στην συνέχεια υλοποιήθηκαν εφαρμογές για την απάντηση των ερωτημάτων της εργασίας

* Στην εφαρμογή 1.py διαβάστηκε η βάση δεδομένων, και ανά 7000 εγγραφές (number\_of\_classes) και για 10 τάξεις (size\_of\_classes) χωρίστηκε το σύνολο των δεδομένων σε μικρότερα τμήματα όπου και αποθηκεύτηκαν ξεχωριστά σε διαφορετικούς φακέλους το καθένα (με όνομα 0 - 9), μέσα στον φάκελο data.
* Με την εφαρμογή 2.py δημιουργείται αρχείο time.json όπου αποθηκεύονται τα χαρακτηρίστηκα των τάξεων που δημιουργήθηκαν. Παρακάτω παρουσιάζονται τα αποτελέσματα σε πίνακα.

|  |  |  |  |
| --- | --- | --- | --- |
| **Class Number** | **From Date** | **To Date** | **Class Length** |
| 1 | 2008-08-01 08:17:57 | 2008-08-21 03:43:57 | 7000 |
| 2 | 2008-08-21 03:48:56 | 2008-08-29 18:17:16 | 7000 |
| 3 | 2008-08-29 18:20:18 | 2008-09-07 19:15:13 | 7000 |
| 4 | 2008-09-07 19:27:37 | 2008-09-15 21:46:23 | 7000 |
| 5 | 2008-09-15 21:47:37 | 2008-09-17 06:36:39 | 7000 |
| 6 | 2008-09-17 06:36:58 | 2008-09-18 22:11:52 | 7000 |
| 7 | 2008-09-18 22:12:35 | 2008-09-20 09:03:56 | 7000 |
| 8 | 2008-09-20 09:04:28 | 2008-09-23 19:06:21 | 7000 |
| 9 | 2008-09-23 19:07:13 | 2008-09-26 01:15:03 | 7000 |
| 10 | 2008-09-26 01:15:25 | 2008-09-29 13:42:31 | 7000 |

* Με την εφαρμογή 4.py υπολογίζονται και αποθηκεύονται οι δείκτες κεντρικότητας για κάθε υπογράφο, και η γραφική απεικόνιση αυτών σε εικόνες .png. Τα αποτελέσματα αποθηκεύονται στον φάκελο του κάθε δικτύου στον φάκελο data. Οι δείκτες που υπολογίστηκαν είναι οι παρακάτω:
  + Degree Centrality 🡪 DC
  + In-Degree Centrality 🡪 IDC
  + Out-Degree Centrality 🡪 ODC
  + Closeness Centrality 🡪 CC
  + Betweenness Centrality 🡪 BC
  + Eigenvector Centrality 🡪 EC
  + Katz Centrality 🡪 KC
* Με την εφαρμογή 5.py υπολογίζονται και αποθηκεύονται για τα ζεύγη δικτύων πχ 0-1, 1-2, 2-3 κτλ

οι κοινές ακμές ή οι νέες ακμές μεταξύ δύο στιγμών ( 🡪 E\_0\_1\_new)

οι παλιές ακμές μεταξύ δύο στιγμών ( 🡪 E\_0\_1\_old)

Κοινοί κόμβοι των δύο στιγμών ( 🡪 V\_0\_1)

Τα αποτελέσματα αποθηκεύονται στον φάκελο data/asteria.

* Με την εφαρμογή 6.py για κάθε ζεύγος νέων Κόμβων υπολογίστηκαν και αποθηκεύτηκαν οι πίνακες ομοιότητας:
* Graph Distance 🡪 GD
* Common Neighbors 🡪 CN
* Jaccard’s Coefficient 🡪 JC
* Adamic / Adar 🡪 A
* Preferential Attachment 🡪 PA

Τα αποτελέσματα αποθηκεύονται στον φάκελο data/similarities.

Οι παραπάνω πίνακες ομοιότητας υπολογίστηκαν για το σύνολο των κόμβων που είναι κοινοί για δύο διαδοχικά υποδίκτυα, δηλαδή το σύνολο , αλλά πάνω στην βάση του συνόλου των ακμών . Πρόκειται για τις ακμές της προηγούμενης χρονικής περιόδου που σχηματίζονται όμως μεταξύ κορυφών που ανήκουν στο κοινό σύνολο των κόμβων .

* Με την εφαρμογή 7\_a\_preparation.py στα δεδομένα του data/similarities από το αμέσως προηγούμενο βήμα προστίθεται η πληροφορία για κάθε ζεύγος ακμών αν είναι παλιά ή νέα ακμή. Βάσει των νέων ακμών θα επαληθευτεί το ποσοστό επιτυχίας των προβλέψεων.

Τα αποτελέσματα αποθηκεύονται στον φάκελο data/predict\_preparation.

* Με την εφαρμογή 7\_b\_success\_limits.py για κάθε υπογράφημα και για κάθε πίνακα ομοιότητας , , και , υπολογίστηκαν τα όρια πάνω από τα οποία μία τιμή του δείκτη ομοιότητας θα θεωρούμε ότι δημιουργεί νέα ακμή. Η λογική που θέτουμε το όριο αυτό είναι ότι οι θεωρητικές νέες ακμές θα είναι 20 φορές περισσότερες σε σύγκριση με τις πραγματικά νέες ακμές.

Τα αποτελέσματα παρουσιάζονται στον παρακάτω πίνακα.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class Name | Similarity | Total New Edges | Percent New Edges (%) | Similarity Value Limit |
| 0\_1 | GD | 2623 | 0.37 | -2.00 |
| 0\_1 | CN | 2623 | 0.37 | 1.00 |
| 0\_1 | JC | 2623 | 0.37 | 0.05 |
| 0\_1 | A | 2623 | 0.37 | 0.42 |
| 0\_1 | PA | 2623 | 0.37 | 400.00 |
| 1\_2 | GD | 2785 | 0.25 | -2.00 |
| 1\_2 | CN | 2785 | 0.25 | 1.00 |
| 1\_2 | JC | 2785 | 0.25 | 0.05 |
| 1\_2 | A | 2785 | 0.25 | 0.33 |
| 1\_2 | PA | 2785 | 0.25 | 304.00 |
| 2\_3 | GD | 2897 | 0.22 | -2.00 |
| 2\_3 | CN | 2897 | 0.22 | 1.00 |
| 2\_3 | JC | 2897 | 0.22 | 0.05 |
| 2\_3 | A | 2897 | 0.22 | 0.31 |
| 2\_3 | PA | 2897 | 0.22 | 240.00 |
| 3\_4 | GD | 787 | 0.07 | -2.00 |
| 3\_4 | CN | 787 | 0.07 | 2.00 |
| 3\_4 | JC | 787 | 0.07 | 0.09 |
| 3\_4 | A | 787 | 0.07 | 0.55 |
| 3\_4 | PA | 787 | 0.07 | 384.00 |
| 4\_5 | GD | 1891 | 0.07 | -2.00 |
| 4\_5 | CN | 1891 | 0.07 | 1.00 |
| 4\_5 | JC | 1891 | 0.07 | 0.04 |
| 4\_5 | A | 1891 | 0.07 | 0.30 |
| 4\_5 | PA | 1891 | 0.07 | 116.00 |
| 5\_6 | GD | 2602 | 0.09 | -3.00 |
| 5\_6 | CN | 2602 | 0.09 | 0.00 |
| 5\_6 | JC | 2602 | 0.09 | 0.00 |
| 5\_6 | A | 2602 | 0.09 | 0.00 |
| 5\_6 | PA | 2602 | 0.09 | 88.00 |
| 6\_7 | GD | 2539 | 0.09 | -3.00 |
| 6\_7 | CN | 2539 | 0.09 | 0.00 |
| 6\_7 | JC | 2539 | 0.09 | 0.00 |
| 6\_7 | A | 2539 | 0.09 | 0.00 |
| 6\_7 | PA | 2539 | 0.09 | 99.00 |
| 7\_8 | GD | 2836 | 0.10 | -2.00 |
| 7\_8 | CN | 2836 | 0.10 | 1.00 |
| 7\_8 | JC | 2836 | 0.10 | 0.03 |
| 7\_8 | A | 2836 | 0.10 | 0.25 |
| 7\_8 | PA | 2836 | 0.10 | 120.00 |
| 8\_9 | GD | 3016 | 0.10 | -3.00 |
| 8\_9 | CN | 3016 | 0.10 | 0.00 |
| 8\_9 | JC | 3016 | 0.10 | 0.00 |
| 8\_9 | A | 3016 | 0.10 | 0.00 |
| 8\_9 | PA | 3016 | 0.10 | 96.00 |

Τα αποτελέσματα αποθηκεύονται στο αρχείο data/limits.json.

* Με την εφαρμογή 7\_c\_predict.py για κάθε υπογράφημα , για κάθε ακμή και για κάθε δείκτη ομοιότητας, βάσει των ορίων που υπολογίστηκαν παραπάνω, υπολογίστηκαν αν μία ακμή θεωρητικά θα εμφανιστεί στο επόμενο στιγμιότυπο του γραφήματος ή όχι.

Τα αποτελέσματα αποθηκεύονται στον φάκελο data/predict.

* Με την εφαρμογή 7\_d\_stats.py για κάθε υπογράφημα υπολογίστηκαν στατιστικά για το πλήθος των θεωρητικών νέων ακμών που προβλέφθηκαν, καθώς και το ποσοστό επιτυχίας πρόβλεψης σε σχέση με τις πραγματικές νέες ακμές. Τα αποτελέσματα παρουσιάζονται στον παρακάτω πίνακα.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Class Name** | **Total new edges** | **Similarity** | **Total  edges** | **Succefully predicted edges** | **Not succefully predicted  edges** |
| 0\_1 | 2623 | GD | 148610 | 913 (1%) | 147697 (99%) |
| 0\_1 | 2623 | CN | 145241 | 1040 (1%) | 144201 (99%) |
| 0\_1 | 2623 | JC | 53089 | 467 (1%) | 52622 (99%) |
| 0\_1 | 2623 | A | 53490 | 588 (1%) | 52902 (99%) |
| 0\_1 | 2623 | PA | 52552 | 580 (1%) | 51972 (99%) |
| 1\_2 | 2785 | GD | 131316 | 845 (1%) | 130471 (99%) |
| 1\_2 | 2785 | CN | 126613 | 978 (1%) | 125635 (99%) |
| 1\_2 | 2785 | JC | 56333 | 448 (1%) | 55885 (99%) |
| 1\_2 | 2785 | A | 56205 | 611 (1%) | 55594 (99%) |
| 1\_2 | 2785 | PA | 55921 | 660 (1%) | 55261 (99%) |
| 2\_3 | 2897 | GD | 118239 | 763 (1%) | 117476 (99%) |
| 2\_3 | 2897 | CN | 113508 | 892 (1%) | 112616 (99%) |
| 2\_3 | 2897 | JC | 60186 | 452 (1%) | 59734 (99%) |
| 2\_3 | 2897 | A | 59137 | 589 (1%) | 58548 (99%) |
| 2\_3 | 2897 | PA | 58239 | 623 (1%) | 57616 (99%) |
| 3\_4 | 787 | GD | 91035 | 174 (0%) | 90861 (100%) |
| 3\_4 | 787 | CN | 15857 | 118 (1%) | 15739 (99%) |
| 3\_4 | 787 | JC | 17194 | 87 (1%) | 17107 (99%) |
| 3\_4 | 787 | A | 15761 | 113 (1%) | 15648 (99%) |
| 3\_4 | 787 | PA | 15780 | 83 (1%) | 15697 (99%) |
| 4\_5 | 1891 | GD | 49193 | 109 (0%) | 49084 (100%) |
| 4\_5 | 1891 | CN | 43811 | 206 (0%) | 43605 (100%) |
| 4\_5 | 1891 | JC | 38305 | 190 (0%) | 38115 (100%) |
| 4\_5 | 1891 | A | 38389 | 157 (0%) | 38232 (100%) |
| 4\_5 | 1891 | PA | 37920 | 98 (0%) | 37822 (100%) |
| 5\_6 | 2602 | GD | 232420 | 437 (0%) | 231983 (100%) |
| 5\_6 | 2602 | CN | 2334784 | 2182 (0%) | 2332602 (100%) |
| 5\_6 | 2602 | JC | 2334784 | 2182 (0%) | 2332602 (100%) |
| 5\_6 | 2602 | A | 2334437 | 2135 (0%) | 2332302 (100%) |
| 5\_6 | 2602 | PA | 54954 | 173 (0%) | 54781 (100%) |
| 6\_7 | 2539 | GD | 250536 | 462 (0%) | 250074 (100%) |
| 6\_7 | 2539 | CN | 2172676 | 2151 (0%) | 2170525 (100%) |
| 6\_7 | 2539 | JC | 2172676 | 2151 (0%) | 2170525 (100%) |
| 6\_7 | 2539 | A | 2172340 | 2120 (0%) | 2170220 (100%) |
| 6\_7 | 2539 | PA | 52078 | 166 (0%) | 51912 (100%) |
| 7\_8 | 2836 | GD | 65286 | 192 (0%) | 65094 (100%) |
| 7\_8 | 2836 | CN | 59355 | 297 (1%) | 59058 (99%) |
| 7\_8 | 2836 | JC | 56817 | 283 (0%) | 56534 (100%) |
| 7\_8 | 2836 | A | 59016 | 262 (0%) | 58754 (100%) |
| 7\_8 | 2836 | PA | 59677 | 228 (0%) | 59449 (100%) |
| 8\_9 | 3016 | GD | 282433 | 765 (0%) | 281668 (100%) |
| 8\_9 | 3016 | CN | 2307361 | 2690 (0%) | 2304671 (100%) |
| 8\_9 | 3016 | JC | 2307361 | 2690 (0%) | 2304671 (100%) |
| 8\_9 | 3016 | A | 2307025 | 2651 (0%) | 2304374 (100%) |
| 8\_9 | 3016 | PA | 60967 | 343 (1%) | 60624 (99%) |

Τα αποτελέσματα αποθηκεύονται στον φάκελο data/predict\_statistics.