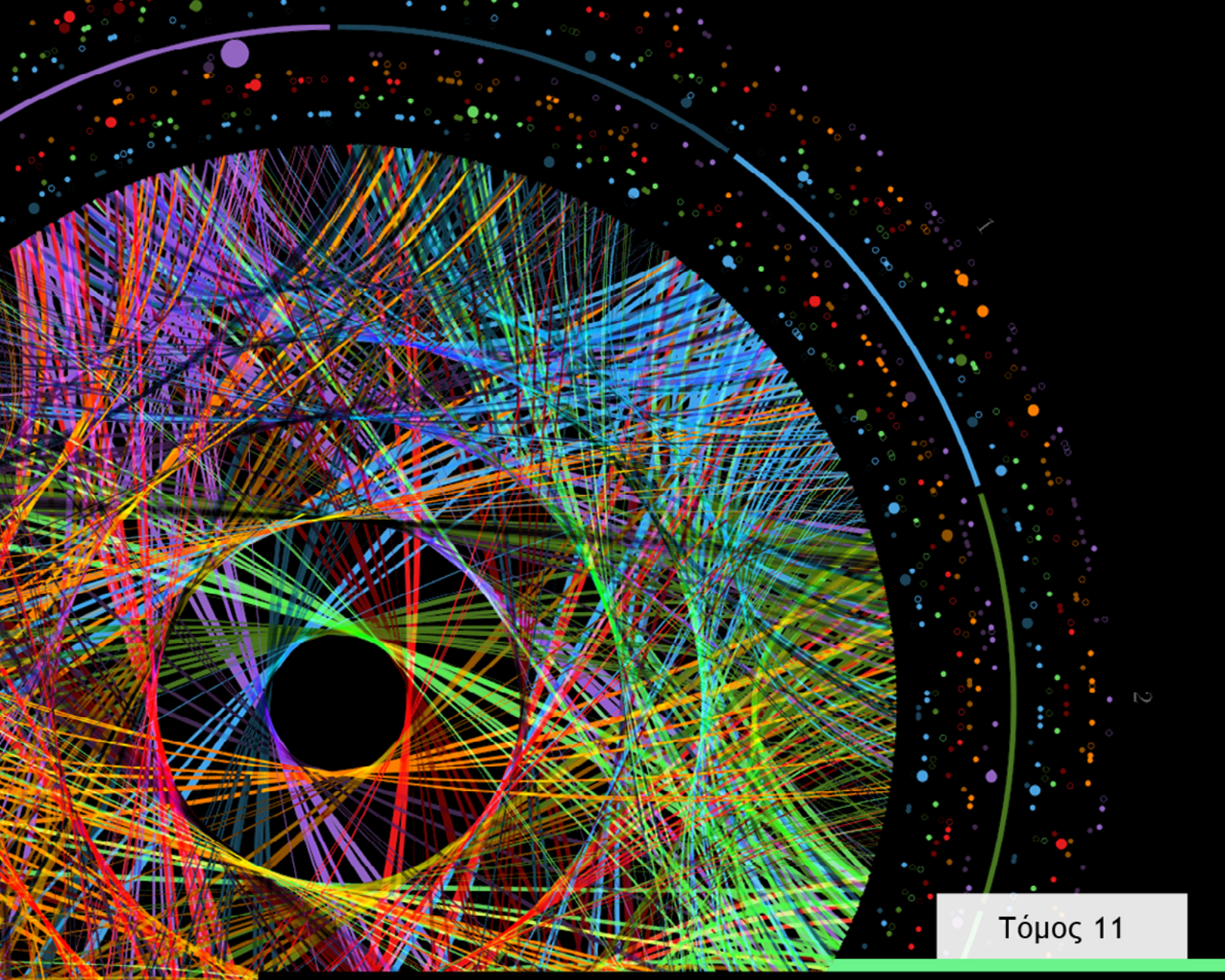


Τόμος 11

ΕΠΙΛΕΓΜΕΝΕΣ

ΠΤΥΧΙΑΚΕΣ & ΔΙΠΛΩΜΑΤΙΚΕΣ ΕΡΓΑΣΙΕΣ

ΑΘΗΝΑ 2014

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
Τμήμα Πληροφορικής και Τηλεπικοινωνιών

ΕΠΙΛΕΓΜΕΝΕΣ ΠΤΥΧΙΑΚΕΣ & ΔΙΠΛΩΜΑΤΙΚΕΣ ΕΡΓΑΣΙΕΣ

Τόμος 11

Εκδίδεται μία φορά το χρόνο από το:

Τμήμα Πληροφορικής και Τηλεπικοινωνιών
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών,
Πανεπιστημιούπολη, 15784 Αθήνα

Τηλ: 210 - 727 5190, Φαξ: 210 - 727 5333
email: *library@di.uoa.gr*, url: *http://www.di.uoa.gr/lib*

Επιμέλεια έκδοσης:

Επιτροπή Ερευνητικών και Αναπτυξιακών Δραστηριοτήτων

Θ. Θεοχάρης (υπεύθυνος έκδοσης), Καθηγητής, Τμήμα Πληροφορικής και Τηλεπικοινωνιών
Η. Μανωλάκος, Αναπληρωτής Καθηγητής, Τμήμα Πληροφορικής και Τηλεπικοινωνιών

Γραφιστική επιμέλεια - Επιμέλεια κειμένων:

Λ. Χαλάτση, Γραφείο Προβολής, Τμήμα Πληροφορικής και Τηλεπικοινωνιών

ISSN 1792-8826

Εξώφυλλο: «Pi Transition Paths» από τη συλλογή Pi Phi e Circos Art του **Martin Krzywinski**.

Ο Martin Krzywinski, επιστήμονας στον τομέα της Βιοπληροφορικής, είναι ο δημιουργός του λογισμικού οπτικοποίησης δεδομένων (data visualization) ***Circos****. Το λογισμικό αυτό προέκυψε από την ανάγκη οπτικοποίησης γονιδιωματικών δεδομένων (genomic data) και πλέον χρησιμοποιείται ευρύτερα σε διάφορους τομείς. Κύριο χαρακτηριστικό του Circos είναι η κυκλική αναπαράσταση των δεδομένων και των συσχετισμών τους, αποδίδοντας τους ένα καλλιτεχνικό αέρα.*

Krzywinski, M. et al. Circos: an Information Aesthetic for Comparative Genomics. Genome Res (2009) 19:1639-1645

Περιεχόμενα

Πρόλογος 5

ΠΤΥΧΙΑΚΕΣ ΕΡΓΑΣΙΕΣ 6

Νικόλαος Σ. Λάριος, Χρήστος Γ. Μητατάκης

Επιλογή Βέλτιστης Διαδρομής χρησιμοποιώντας
Καταγραφή Κυκλοφοριακών Δεδομένων μέσω Κινητών Συσκευών 7

Μαρία - Άννα Γ. Περιδέλη

Προηγμένα Ευρετικά Διαχώρισης Πεδίων Τιμών
σε Προβλήματα Ικανοποίησης Περιορισμών 22

Εμμανουήλ Γ. Τζαγκαράκης

Νευρωνικά Δίκτυα και Αλγόριθμοι Εκπαίδευσης
για Κατηγοριοποίηση Κειμένου 35

ΔΙΠΛΩΜΑΤΙΚΕΣ ΕΡΓΑΣΙΕΣ 49

Blerina P. Lika

A Novel Approach for alleviating the Cold Start Problem
in Recommender Systems 50

Georgios P. Nomikos

Point centrality indices
and ISP network vulnerability.....63

Βασίλειος Α. Τσιρώνης

Μελέτη Μηχανισμού Διασφάλισης Συνέπειας Εξυπηρετητών
Κρυφής Μνήμης Παγκόσμιου Ιστού, με Εφαρμογή της
Θεωρίας Βέλτιστης Παύσης.....78

Πρόλογος

Ο τόμος αυτός περιλαμβάνει περιλήψεις επιλεγμένων διπλωματικών και πτυχιακών εργασιών που εκπονήθηκαν στο Τμήμα Πληροφορικής και Τηλεπικοινωνιών του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών κατά το διάστημα **01/01/2013 - 31/12/2013**. Πρόκειται για τον **11ο τόμο** στη σειρά αυτή. Στόχος του θεσμού είναι η ενθάρρυνση της δημιουργικής προσπάθειας και η προβολή των πρωτότυπων εργασιών των φοιτητών του Τμήματος.

Η έκδοση αυτή είναι ψηφιακή και έχει δικό της ISSN. Αναρτάται στην επίσημη ιστοσελίδα του Τμήματος και έτσι, εκτός από τη μείωση της δαπάνης κατά την τρέχουσα περίοδο οικονομικής κρίσης, έχει και μεγαλύτερη προσβασιμότητα. Για το στόχο αυτό, σημαντική ήταν η συμβολή της Λήδας Χαλάτση που επιμελήθηκε και φέτος την ψηφιακή έκδοση και πέτυχε μια ελκυστική ποιότητα παρουσίασης, ενώ βελτίωσε και την ομοιογένεια των κειμένων.

Η στάθμη των επιλεγμένων εργασιών είναι υψηλή και κάποιες από αυτές έχουν είτε δημοσιευθεί είτε υποβληθεί για δημοσίευση.

Θα θέλαμε να ευχαριστήσουμε τους φοιτητές για το χρόνο που αφιέρωσαν για να παρουσιάσουν τη δουλειά τους στα πλαίσια αυτού του θεσμού και να τους συγχαρούμε για την ποιότητα των εργασιών τους. Ελπίζουμε η διαδικασία αυτή να προσέφερε και στους ίδιους μια εμπειρία που θα τους βοηθήσει στη συνέχεια των σπουδών τους ή της επαγγελματικής τους σταδιοδρομίας.

Η Επιτροπή Ερευνητικών και Αναπτυξιακών Δραστηριοτήτων

Θ. Θεοχάρης (υπεύθυνος έκδοσης), Η. Μανωλάκος

Αθήνα, Ιούλιος 2014

ΠΤΥΧΙΑΚΕΣ ΕΡΓΑΣΙΕΣ

Νικόλαος Σ. Λάριος
nlarios@di.uoa.gr

Χρήστος Γ. Μητατάκης
cmitatakis@di.uoa.gr

Επιλογή Βέλτιστης Διαδρομής χρησιμοποιώντας Καταγραφή Κυκλοφοριακών Δεδομένων μέσω Κινητών Συσκευών

Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, Τμήμα Πληροφορικής και Τηλεπικοινωνιών
Πανεπιστημιούπολη, Ιλίσια, 15784, Αθήνα, Ελλάς

Περίληψη

Η κυκλοφοριακή συμφόρηση των πόλεων σε συνδυασμό με τους σύγχρονους ρυθμούς ζωής καθιστούν αναγκαία την ύπαρξη ενός εργαλείου που θα συμβάλει στη βελτίωση των μεταφορών, παρέχοντας στους χρήστες του τη δυνατότητα εύρεσης της γρηγορότερης διαδρομής για τον εκάστοτε προορισμό τους. Στο παρόν έγγραφο παρουσιάζουμε την εφαρμογή Smart Driver (για κινητές συσκευές πχ. smartphones, tablets με λειτουργικό σύστημα Android) μέσω της οποίας ο χρήστης έχει τη δυνατότητα εύρεσης της βέλτιστης χρονικά διαδρομής για κάποιο προορισμό μαζί με τον προβλεπόμενο χρόνο εκτέλεσης της. Βασικός στόχος της εφαρμογής είναι η βελτίωση των αποτελεσμάτων του Google Directions API, στα οποία ο υπολογισμός του χρόνου μιας διαδρομής δεν βασίζεται σε πραγματικά δεδομένα κίνησης. Ο κορμός της εφαρμογής αποτελείται από δύο κομμάτια, την καταγραφή κυκλοφοριακών δεδομένων και την εύρεση της βέλτιστης χρονικά διαδρομής. Η καταγραφή των κυκλοφοριακών δεδομένων γίνεται με χρήση του GPS των κινητών συσκευών. Η εύρεση της βέλτιστης χρονικά διαδρομής πραγματοποιείται με την αξιοποίηση των πραγματικών δεδομένων που συλλέγονται από την εφαρμογή, σε συνδυασμό με τα αποτελέσματα του Google Directions API. Η εφαρμογή εκμεταλλεύεται την ύπαρξη μοντέλων κίνησης που βασίζονται στα χαρακτηριστικά κάθε ημέρας και ώρας. Η εφαρμογή είναι φιλική προς το χρήστη μέσω κατάλληλης γραφικής διεπαφής.

Λέξεις κλειδιά: Εξόρυξη δεδομένων, εφαρμογή android, βέλτιστη διαδρομή, καταγραφή κυκλοφοριακών δεδομένων, γεωγραφικά σημεία.

Επιβλέπων

Δημήτριος Γουνόπουλος, Καθηγητής

1. Εισαγωγή

Ένα βασικό πρόβλημα με το οποίο έχει ασχοληθεί η επιστήμη της πληροφορικής είναι η εύρεση της βέλτιστης διαδρομής μεταξύ δύο σημείων. Η σχεδίαση μιας διαδρομής βρίσκει πρακτική εφαρμογή στην πλοήγηση οχημάτων πάνω σε ένα οδικό δίκτυο. Το πρόβλημα αυτό αναφέρεται συχνά ως πρόβλημα εύρεσης συντομότερης διαδρομής (Shortest path problem-SP). Για την επίλυση του, έχουν αναπτυχθεί πολλοί αλγόριθμοι οι οποίοι το αντιμετωπίζουν με τεράστια επιτυχία. Αυτοί οι αλγόριθμοι έχουν ενσωματωθεί σε εφαρμογές που εκατομμύρια χρήστες τις χρησιμοποιούν καθημερινά μέσω διαδικτύου, ειδικών κινητών συσκευών (GPS destinators) [1] και πιο πρόσφατα, μέσω σύγχρονων κινητών συσκευών, δηλαδή smartphones και tablets.

Οι αλγόριθμοι των εφαρμογών αυτών χρησιμοποιούν παραμέτρους, όπως το συνολικό μήκος μιας διαδρομής και τα όρια ταχύτητας των οδών για να μπορέσουν να προσφέρουν στους χρήστες τους τη γρηγορότερη διαδρομή. Τα τελευταία χρόνια πολλές εφαρμογές προκειμένου να έχουν αποτελέσματα που ανταποκρίνονται περισσότερο στην πραγματικότητα, αξιοποιούν πραγματικά κυκλοφοριακά δεδομένα, προβλέποντας με ακόμα μεγαλύτερη ακρίβεια το χρόνο διεξαγωγής μιας διαδρομής.

Η δημοφιλέστερη εφαρμογή πλοήγησης λόγω των υπηρεσιών που προσφέρει, αλλά και λόγω της δωρεάν διάθεσης της μέσω διαδικτύου και για κινητές συσκευές με λειτουργικό σύστημα Android, είναι η εφαρμογή Google maps της Google [3]. Η εφαρμογή αυτή έχει τη δυνατότητα σε ελάχιστο χρόνο να

επιστρέψει στο χρήστη μέχρι και τρεις πιθανές διαδρομές, από ένα σημείο Α μέχρι ένα σημείο Β, μαζί με τον πιθανό χρόνο διεξαγωγής της διαδρομής. Αυτός ο χρόνος, στη χώρα μας υπολογίζεται μόνο από τα όρια ταχύτητας των οδών της κάθε διαδρομής, δηλαδή δεν χρησιμοποιούνται πραγματικά κυκλοφοριακά δεδομένα. Επομένως, ο πιθανός χρόνος διεξαγωγής μιας διαδρομής απέχει συχνά πολύ από τη πραγματικότητα, μειώνοντας έτσι την αξιοπιστία της εφαρμογής.

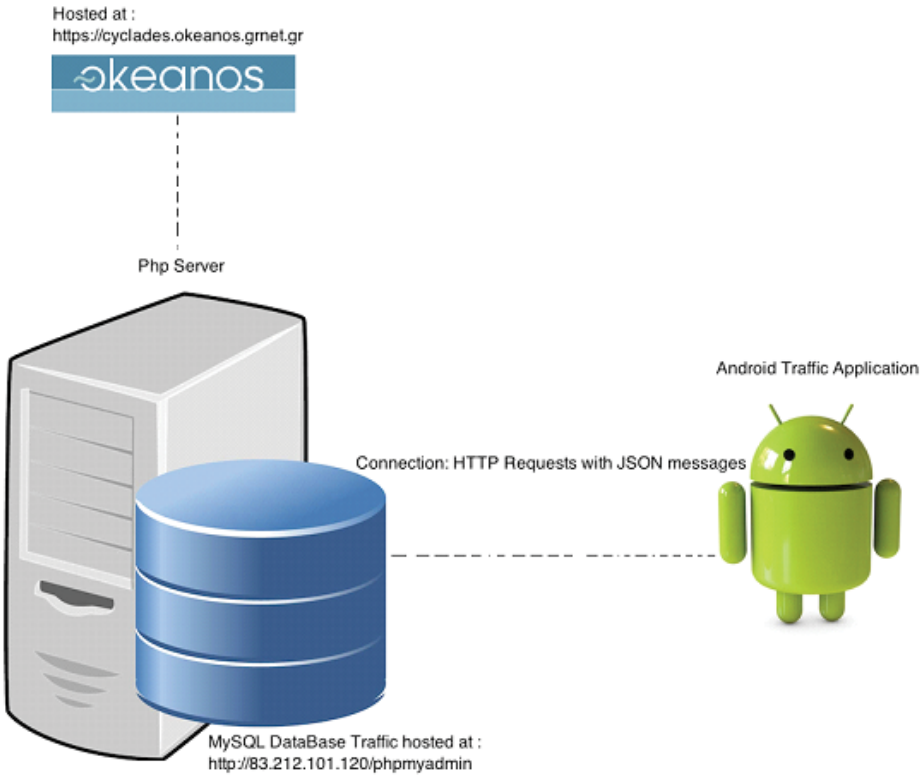
Στα πλαίσια αυτής της πτυχιακής εργασίας, αναπτύχθηκε εφαρμογή για κινητές συσκευές με λειτουργικό σύστημα Android 2.3 ή νεότερη έκδοση, η οποία έχει ως στόχο τη βελτίωση των αποτελεσμάτων της εφαρμογής Google maps, με χρήση πραγματικών κυκλοφοριακών δεδομένων τα οποία αυτή καταγράφει. Η εφαρμογή εκμεταλλεύεται το γεγονός ότι η κίνηση των δρόμων σχετίζεται άμεσα με κάθε ημέρα και ώρα της εβδομάδας. Δηλαδή, η εφαρμογή καταγράφει κυκλοφοριακά δεδομένα με χρήση εργαλείων της συσκευής, όπως το GPS και το ασύρματο δίκτυο 3G, και με βάση αυτά επιστρέφει στο χρήστη της τη βέλτιστη χρονικά διαδρομή μεταξύ δύο σημείων της επιλογής του. Η βέλτιστη διαδρομή που επιστρέφει στο χρήστη η εφαρμογή, είναι η γρηγορότερη διαδρομή από τις επιλογές του Google Maps, με βάση τα κυκλοφοριακά δεδομένα που έχουν συλλεχθεί για τη συγκεκριμένη μέρα και ώρα.

Επιπλέον, ο προβλεπόμενος χρόνος ολοκλήρωσης της διαδρομής που επιστρέφεται στο χρήστη είναι πιο κοντά στη πραγματικότητα, καθώς υπολογίζεται με βάση τα κυκλοφοριακά δεδομένα που έχουν συλλεχθεί. Η εφαρμογή αποτελείται από τη διεπαφή της κινητής συσκευής Android και έναν εξυπηρετητή (Server). Η γλώσσα προγραμματισμού που επιλέχθηκε για τον server της εφαρμογής είναι η PHP και ο κώδικας αναπτύχθηκε και δοκιμάστηκε σε virtual machine Ubuntu server του συστήματος του Okeanos. Η ανάπτυξη της διεπαφής της κινητής συσκευής πραγματοποιήθηκε σε γλώσσα Java και περιβάλλον ανάπτυξης Eclipse με Android Software Development Kit και μπορεί να τρέξει σε οποιαδήποτε Android συσκευή που διαθέτει λειτουργικό σύστημα Android 2.3 ή νεότερη έκδοση.

2. Περιγραφή εφαρμογής Android

2.1. Σχήμα Αρχιτεκτονικής της Εφαρμογής

Το σχήμα της αρχιτεκτονικής της εφαρμογής είναι το εξής:



Εικόνα 1: Σχήμα αρχιτεκτονικής της εφαρμογής

Όπως φαίνεται αποτελείται από 3 μέρη, την εφαρμογή του εξυπηρετητή, μία βάση δεδομένων και την εφαρμογή της κινητής συσκευής Android.

2.2. Παρουσίαση της εφαρμογής Android

2.2.1. Γενικές πληροφορίες

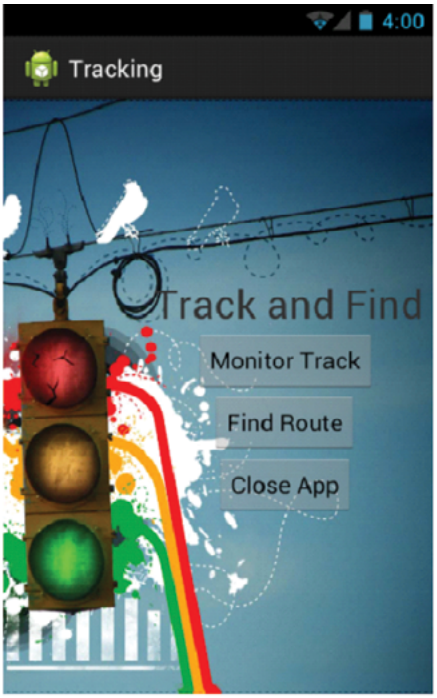
Η εφαρμογή Android αναπτύχθηκε με χρήση του Android SDK, το οποίο περιλαμβάνει το περιβάλλον ανάπτυξης εφαρμογών Eclipse IDE με ενσωματωμένο ADT (Android Development Tools) και το οποίο παρέχει όλες τις βιβλιοθήκες API και τα εργαλεία που είναι απαραίτητα για την κατασκευή, τη δοκιμή και τις εφαρμογές εντοπισμού σφαλμάτων για το Android. Το Android SDK είναι συμβατό με τα λειτουργικά συστήματα Windows και Mac OS. Οι γλώσσες προγραμματισμού που χρησιμοποιήθηκαν για την ανάπτυξη της εφαρμογής ήταν JAVA και xml. Οι δοκιμές κατά την ανάπτυξη και αποσφαλμάτωση της εφαρμογής έγιναν στη συσκευή Samsung Nexus S και σε λειτουργικό σύστημα

Android version 4.0.4.

Η εφαρμογή αποτελείται από δύο βασικά κομμάτια. Αυτό της καταγραφής των κυκλοφοριακών δεδομένων από το χρήστη και αυτό της εύρεσης της βέλτιστης διαδρομής μεταξύ δύο σημείων που επιλέγει ο χρήστης. Για τη χρήση της εφαρμογής είναι απαραίτητο η εφαρμογή να έχει συνεχή σύνδεση στο διαδίκτυο είτε μέσω κάποιου δικτύου wifi είτε μέσω ασύρματου δικτύου 3G-4G. Στη συνέχεια ακολουθεί αναλυτική παρουσίαση της εφαρμογής και οδηγίες για τη χρήση της.

2.2.2. Κεντρικό μενού εφαρμογής.

Το κεντρικό μενού της εφαρμογής έχει την εξής μορφή:



Εικόνα 2: Κεντρικό μενού εφαρμογής

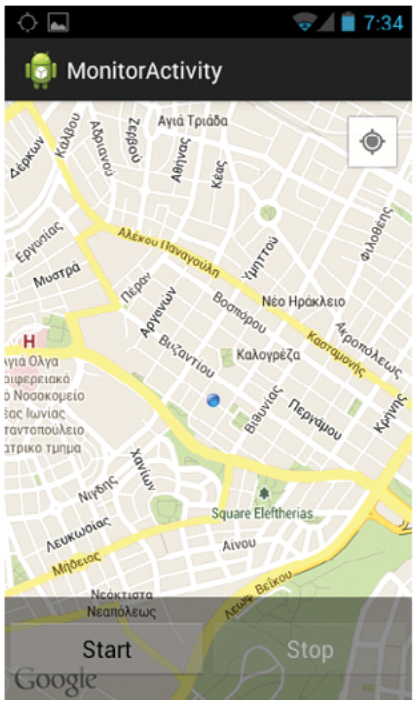
Όπως φαίνεται, το κεντρικό μενού αποτελείται από μια επικεφαλίδα, μία εικόνα φόντου και τρία κουμπιά τα οποία δίνουν στον χρήστη τις εξής επιλογές:

- Monitor Track (Καταγραφή διαδρομής): Επιλέγεται από το χρήστη όταν θέλει να καταγράψει τα κυκλοφοριακά δεδομένα της διαδρομής του. Οδηγείται στην οθόνη καταγραφής διαδρομής που περιγράφεται παρακάτω.

- Find Route (Εύρεση Διαδρομής): Επιλέγεται από το χρήστη όταν θέλει να βρει τη βέλτιστη διαδρομή μεταξύ δύο σημείων. Οδηγείται στην οθόνη εύρεσης διαδρομής που περιγράφεται στη συνέχεια.
- Close App (Κλείσιμο Εφαρμογής): Ο χρήστης εξέρχεται από την εφαρμογή σταματώντας τη λειτουργία της.

2.2.3. Οθόνη καταγραφής διαδρομής (MonitorActivity)

Εφόσον ο χρήστης επιλέξει στο κεντρικό μενού την επιλογή Monitor Track οδηγείται στην οθόνη καταγραφής διαδρομής:

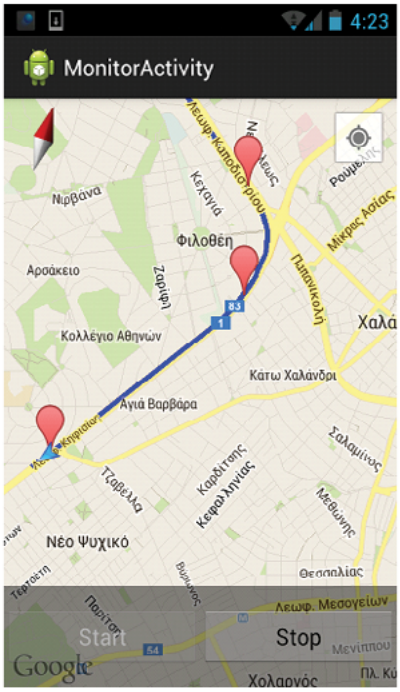


Εικόνα 3: Οθόνη καταγραφής κυκλοφοριακών δεδομένων εφαρμογής

Η οθόνη αυτή, όπως φαίνεται, αποτελείται από έναν χάρτη και δύο κουμπιά (Start, Stop). Ο χάρτης χρησιμοποιείται από την εφαρμογή μέσω του Google Maps Android v2 API που περιγράφεται στο κεφάλαιο 2. Κατά την είσοδο του χρήστη σε αυτή την οθόνη μπορεί να επιλέξει μονάχα την επιλογή “Start” (ή να επιστρέψει στην προηγούμενη οθόνη με το πλήκτρο “πίσω” της συσκευής). Με την επιλογή “Start” ξεκινάει η καταγραφή της διαδρομής του χρήστη. Εφόσον έχει πατηθεί το “Start” δεν είναι πια επιλέξιμο, ενώ αντίθετα η επιλογή “Stop” γίνεται επιλέξιμη. Η τοποθεσία του χρήστη εμφανίζεται στο χάρτη με ένα μπλε στίγμα όταν είναι ακίνητος και με ένα μπλε βέλος με τη

κατεύθυνση του όταν κινείται. Όταν η συσκευή κινείται στον χάρτη εμφανίζεται με μπλε ίχνος η τροχιά που ακολουθείται. Επιπλέον όταν η συσκευή περνάει κοντά από κάποιο σημείο ενδιαφέροντος (τα οποία περιγράφονται αναλυτικά παρακάτω και σε αυτά γίνεται η καταγραφή κυκλοφοριακών δεδομένων) εμφανίζεται στο σημείο αυτό ένας κόκκινος δείκτης (“Marker”).

Ένα στιγμιότυπο της εφαρμογής κατά την καταγραφή κυκλοφοριακών δεδομένων και ενώ η συσκευή βρίσκεται σε κίνηση είναι το εξής:



Εικόνα 4: Οθόνη καταγραφής κυκλοφοριακών δεδομένων κατά την καταγραφή

Εφόσον ο χρήστης επιθυμεί να σταματήσει την καταγραφή δεδομένων, τότε πρέπει είτε να πατήσει το κουμπί “Stop” είτε να πατήσει το πλήκτρο “πίσω” της συσκευής. Όταν ο χρήστης σταματήσει την καταγραφή δεδομένων, εμφανίζεται μια ροδέλα φόρτωσης όσο η εφαρμογή στέλνει στον server τα δεδομένα που έχει συλλέξει και δεν έχουν σταλεί ακόμα. Έπειτα η επιλογή “Start” είναι ξανά διαθέσιμη για επιλογή.

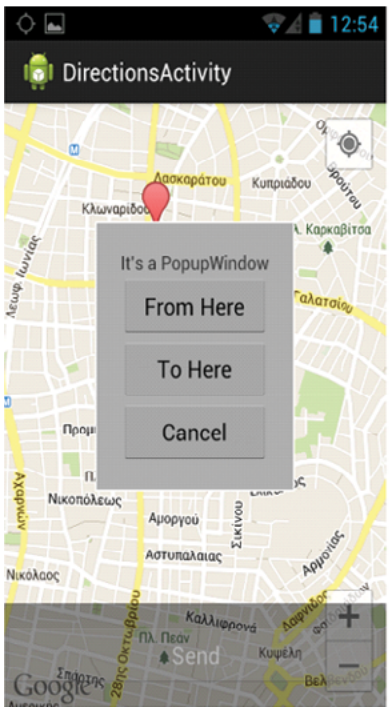
2.2.4. Οθόνη εύρεσης διαδρομής (DirectionsActivity)

Εφόσον ο χρήστης επιλέξει την επιλογή “Find Route”, οδηγείται στην οθόνη

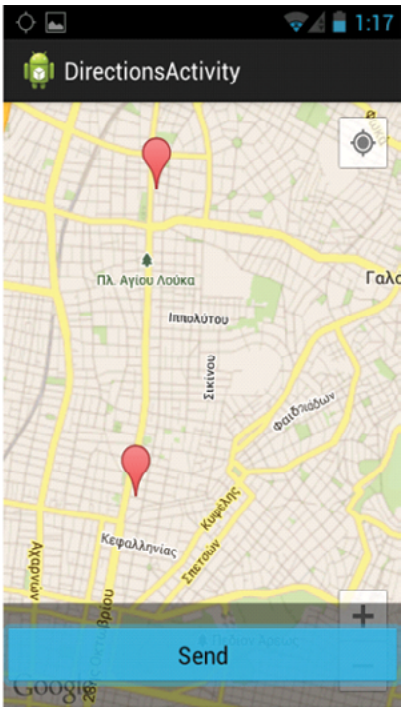
επιλογής των δύο σημείων για τα οποία θέλει να βρει τη βέλτιστη διαδρομή. Η οθόνη αυτή αποτελείται από το χάρτη του Google Maps, όπως αυτός προσφέρεται από το Google Maps API, το στίγμα της συσκευής πάνω στο χάρτη και ένα κουμπί “Send” το οποίο όμως δεν είναι επιλέξιμο:

Στη συνέχεια, για να επιλέξει ο χρήστης τα δύο σημεία μεταξύ των οποίων θέλει να βρει τη διαδρομή, ακολουθεί τα βήματα:

- Αρχικά ακουμπάει την οθόνη στο 1ο σημείο που τον ενδιαφέρει ώσπου να εμφανιστεί ένας κόκκινος δείκτης (Marker) στο σημείο και ένα παράθυρο με της επιλογές “From Here”, “To Here” και “Cancel”.
- Σε περίπτωση που το σημείο αυτό είναι το σημείο έναρξης της διαδρομής, πρέπει να επιλέξει την επιλογή “From Here”.
- Στην περίπτωση όπου το σημείο που επέλεξε είναι το σημείο τερματισμού της διαδρομής, τότε πρέπει να επιλέξει την επιλογή “To Here”.
- Σε περίπτωση λάθους μπορεί να επιλέξει “Cancel” και ο δείκτης διαγράφεται.



Εικόνα 5: Επιλογή σημείου



Εικόνα 6: Αποστολή ερωτήματος εύρεσης διαδρομής

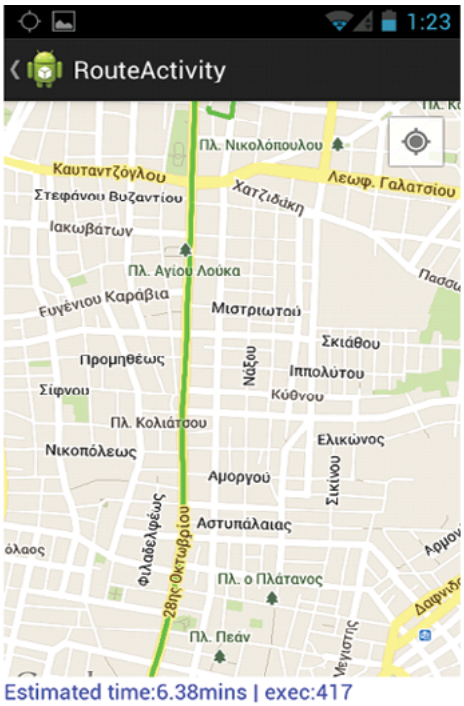
- Εφόσον ο χρήστης επιλέξει δύο σημεία, ένα έναρξης διαδρομής και ένα τερματισμού, η επιλογή “Send” στο κάτω μέρος της οθόνης γίνεται επιλέξιμη. Εφόσον ο χρήστης επιλέξει “Send”, στέλνεται το ερώτημα της διαδρομής

στον Server και εφόσον δεχτεί απάντηση οδηγεί το χρήστη στην επόμενη οθόνη παρουσίασης της βέλτιστης διαδρομής.

- Εάν ο χρήστης έχει επιλέξει κάποιο λάθος σημείο μπορεί να το διορθώσει ακουμπώντας τον κόκκινο δείκτη του σημείου και επιλέγοντας “Delete” στο παράθυρο που θα εμφανιστεί.

2.2.5. Οθόνη παρουσίασης βέλτιστης διαδρομής (RouteActivity)

Αφότου ο χρήστης έχει ζητήσει τη βέλτιστη διαδρομή μεταξύ δύο σημείων, οδηγείται στην οθόνη παρουσίασης βέλτιστης διαδρομής, η οποία αποτελείται από το γνωστό χάρτη, ένα ίχνος της επιλεγμένης βέλτιστης διαδρομής και ένα άσπρο πεδίο στο κάτω μέρος της οθόνης στο οποίο αναγράφεται ο προβλεπόμενος χρόνος ολοκλήρωσης της διαδρομής, όπως αυτός έχει υπολογιστεί με βάση τα δεδομένα που έχει διαθέσιμα η εφαρμογή.



Εικόνα 7: Προβολή βέλτιστης διαδρομής χωρίς κυκλοφοριακή συμφόρηση

Το χρώμα του ίχνους της βέλτιστης διαδρομής είναι πράσινο για τα τμήματα της διαδρομής για τα οποία δεν υπάρχει καθόλου κυκλοφοριακή συμφόρηση, με γκρι χρώμα απεικονίζεται η μέτρια συμφόρηση και με κόκκινο χρώμα εμφανίζονται τα τμήματα της διαδρομής με μεγάλη κυκλοφοριακή συμφόρηση, σύμφωνα πάντα με τα δεδομένα που έχουν συλλεχθεί από την εφαρμογή.

Επιπλέον, όταν ο χρήστης βρίσκεται σε αυτή την οθόνη γίνεται καταγραφή δεδομένων όπως και στην οθόνη καταγραφής δεδομένων. Δηλαδή, παρουσιάζεται η πορεία που ακολουθεί με μπλε ίχνος και τα σημεία ενδιαφέροντος από τα οποία περνάει εμφανίζονται με κόκκινους δείκτες πάνω στον χάρτη.

3. Περιγραφή εξυπηρετητή εφαρμογής

3.1. Παρουσίαση του Server της εφαρμογής και της βάσης δεδομένων

3.1.1. Γενικές Πληροφορίες

Για τις ανάγκες της εφαρμογής, δημιουργήθηκε ένας Ubuntu Server και μια βάση δεδομένων MySQL σε ένα virtual machine στο σύστημα Okeanos του Εθνικού Μετσόβιου Πολυτεχνείου (*https://okeanos.grnet.gr/home/*), το οποίο δίνει τη δυνατότητα δημιουργίας δωρεάν λογαριασμού για φοιτητές. Ο Server αποτελεί web service με χρήση της γλώσσας PHP καθώς και του Apache Server.

Ο Server υλοποιεί τις εξής λειτουργίες:

- Δέχεται ένα αίτημα από την εφαρμογή Android με όρισμα τις συντεταγμένες που βρίσκεται εκείνη τη στιγμή ο χρήστης και μέσω http request στη google maps βρίσκει σε ποιο δρόμο βρίσκεται.
- Δέχεται ένα αίτημα από την εφαρμογή Android με όρισμα το δρόμο που βρίσκεται εκείνη τη στιγμή ο χρήστης, ελέγχει αν υπάρχει στη βάση και αν υπάρχει της επιστρέφει τα σημεία ενδιαφέροντος του εν λόγω δρόμου.
- Δέχεται δύο σημεία ενδιαφέροντος με τις συντεταγμένες τους, το συνολικό χρόνο ανάμεσα σε αυτά, την ταχύτητα που πέρασε ο χρήστης από αυτά και το timestamp και τα αποθηκεύει στον πίνακα TrafficData, καθώς και στον αντίστοιχο συνολικό πίνακα της ημέρας.
- Δέχεται ένα αίτημα από την εφαρμογή Android με όρισμα τις συντεταγμένες που βρίσκεται εκείνη τη στιγμή ο χρήστης και τις συντεταγμένες προορισμού και μέσω http request στη Google Maps βρίσκει τις προτεινόμενες διαδρομές και επιλέγει τη βέλτιστη χρονικά διαδρομή βάση των κυκλοφοριακών δεδομένων.

Ουσιαστικά ο Server είναι υπεύθυνος για όλες τις αιτήσεις στο Google Maps API

(η οποία γίνεται με μηνύματα JSON), καθώς και για τη διαχείριση της βάσης δεδομένων.

3.1.2. Εξυπηρέτηση των Αιτημάτων της εφαρμογής Android από τον Server

Ο Server είναι υπεύθυνος για την εξυπηρέτηση διάφορων αιτήσεων της εφαρμογής Android σε αυτόν και την αποστολή των κατάλληλων δεδομένων σε αυτή.

Συγκεκριμένα έχουμε:

- Περίπτωση της καταγραφής δεδομένων: Ο Server δέχεται τα δεδομένα που του στέλνει η εφαρμογή Android και τα αποθηκεύει στον κατάλληλο πίνακα της βάσης δεδομένων. Παραπάνω λεπτομέρειες θα αναφερθούν στο επόμενο υποκεφάλαιο του εν λόγω κεφαλαίου.
- Περίπτωση αποστολής γρηγορότερης διαδρομής: Ο Server αποδέχεται από κάποιον χρήστη μια ερώτηση για μια διαδρομή η οποία περιλαμβάνει την τρέχουσα τοποθεσία του (longtitude, latitude) και την τοποθεσία στην οποία θέλει να κατευθυνθεί. Ο Server στο σημείο αυτό με τη χρήση του Google Directions API [2] βρίσκει 3 διαδρομές. Επιπλέον, κρατάει σε διαφορετικές μεταβλητές τον εκτιμώμενο χρόνο που επιστρέφει το Google Maps Api για την εκάστοτε διαδρομή. Ο εκτιμώμενος χρόνος της διαδρομής θα χρησιμοποιηθεί σαν παράγοντας συμφόρησης. Θα συγκριθεί με το χρόνο εκτέλεσης μιας παρόμοιας τροχιάς σύμφωνα με τα δεδομένα που έχουμε συλλέξει, λαμβάνοντας υπόψη την ώρα και τη μέρα που έχει συλλεχθεί η αντίστοιχη τροχιά. Ο αλγόριθμος, αρχικά, ελέγχει στη βάση δεδομένων αν υπάρχουν παρόμοιες τροχιές τη συγκεκριμένη ώρα και μέρα. Αν υπάρχουν και ο παράγοντας συμφόρησης είναι μικρός, τότε η συντομότερη από αυτές τις τροχιές προτείνεται στο χρήστη. Αν ο παράγοντας συμφόρησης είναι μεγάλος, τότε επιλέγουμε τη διαδρομή με το μικρότερο συνολικό χρόνο. Τέλος, σε περίπτωση που δεν υπάρχει στη βάση μας κάποια παρόμοια διαδρομή επιλέγεται απλά η κοντινότερη διαδρομή που δίνει το Google Maps API. Σε κάθε περίπτωση αποστέλλεται η προτεινόμενη τελική διαδρομή στο χρήστη με τη μορφή μηνύματος JSON.

4. Συμπεράσματα

Το πρόβλημα εύρεσης βέλτιστης διαδρομής μεταξύ δύο σημείων πάνω σε ένα χάρτη έχει απασχολήσει ιδιαίτερα την επιστήμη της πληροφορικής. Τα βήματα για τη λύση του ήταν η μοντελοποίηση του προβλήματος σε ένα γράφο, έτσι ώστε να μπορεί να λυθεί αλγοριθμικά και στη συνέχεια η δημιουργία του κατάλληλου αλγορίθμου. Ο βασικότερος αλγόριθμος ο οποίος χρησιμοποιείται μέχρι σήμερα είτε αυτούσιος, είτε με βελτιστοποιήσεις είναι ο αλγόριθμος του Dijkstra[4].

Με τις τεχνολογικές εξελίξεις και με τη δημιουργία ηλεκτρονικών χαρτών, αυτοί οι αλγόριθμοι και η μοντελοποίηση του προβλήματος βρήκαν μεγάλη πρακτική εφαρμογή. Οι εφαρμογές οι οποίες έλυναν το πρόβλημα αυτό γνώρισαν μεγάλη εμπορική επιτυχία. Ειδικότερα με τον ερχομό του GPS, αυτές οι εφαρμογές μετεξελίχθηκαν σε εφαρμογές πλοήγησης, οι οποίες έτρεχαν σε αυτόνομες φορητές συσκευές που έγιναν γρήγορα δημοφιλείς στους οδηγούς σε ολόκληρο τον κόσμο. Με τον ερχομό των έξυπνων κινητών συσκευών, δηλαδή των smartphones και των tablets, και με τις τεράστιες δυνατότητες που αυτές προσφέρουν ανοίχτηκαν νέοι δρόμοι για τη δημιουργία εφαρμογών οι οποίες θα προσφέρουν στους χρήστες τους ακόμα καλύτερες υπηρεσίες πλοήγησης. Οι μεγάλες εταιρείες συνέχισαν να ασχολούνται με το πρόβλημα, προσπαθώντας να προσφέρουν καλύτερα αποτελέσματα, αφού η βέλτιστη διαδρομή σε έναν οδικό χάρτη εξαρτάται από πολλούς παράγοντες και όχι μόνο από το μήκος της διαδρομής. Οι παράγοντες αυτοί μπορεί να είναι η ποιότητα των δρόμων που επιλέγονται, δηλαδή η ασφάλεια τους και η μέση ταχύτητα με την οποία κινούνται τα οχήματα σε αυτόν και η κυκλοφοριακή συμφόρηση. Με λίγα λόγια, τον οδηγό τον ενδιαφέρει κυρίως ποια είναι η πιο γρήγορη διαδρομή, καθώς αυτή θα είναι και η οικονομικότερη. Επομένως, μια σημαντική βελτιστοποίηση των εφαρμογών πλοήγησης είναι ο συνυπολογισμός της κυκλοφοριακής συμφόρησης μιας διαδρομής. Ιδιαίτερα σε μια πόλη σαν αυτή της Αθήνας, οι κυκλοφοριακές συνθήκες είναι ίσως ο βασικότερος παράγοντας ο οποίος μεταβάλει το χρόνο εκτέλεσης μιας διαδρομής. Η εφαρμογή που αναπτύχθηκε και περιγράφεται σε αυτό το έγγραφο, έχει σκοπό να συμπεριλάβει αυτή τη βελτιστοποίηση στην δημοφιλέστερη εφαρμογή πλοήγησης, το Google Maps.

Η φορητότητα των κινητών συσκευών σε συνδυασμό με τις δυνατότητες τους που όλο και αυξάνονται, είναι ένα χρήσιμο εργαλείο που μπορεί να βελτιώσει σε πολλούς τομείς τη καθημερινότητα μας. Οι δυνατότητες τους

πολλαπλασιάζονται αν λειτουργούν κατανεμημένα, δηλαδή όταν πολλές κινητές συσκευές τρέχουν την ίδια εφαρμογή με έναν κοινό στόχο. Συχνά, όσο περισσότερες κινητές συσκευές τρέχουν μία εφαρμογή τόσο το τελικό της αποτέλεσμα είναι καλύτερο. Η εφαρμογή που περιγράφεται στο παρόν εκμεταλλεύεται αυτή τη δυνατότητα των κινητών συσκευών, καθώς ο σκοπός της είναι η συλλογή κυκλοφοριακών δεδομένων από κινητές συσκευές πολλών χρηστών.

Η εγκυρότητα των αποτελεσμάτων της εφαρμογής εξαρτάται από τη ποσότητα των κυκλοφοριακών δεδομένων τα οποία έχουμε στη διάθεση μας (δηλαδή του μεγέθους του Dataset). Επομένως, βασικός παράγοντας επιτυχίας είναι ο αριθμός των χρηστών για δειγματοληψία. Όσο μεγαλύτερος είναι ο αριθμός των χρηστών της εφαρμογής τόσο ακριβέστερα είναι τα τελικά αποτελέσματα και μπορούν να αποφευχθούν στατιστικά σφάλματα ακραίων τιμών. Επιπλέον, για την καλύτερη λειτουργία της εφαρμογής ιδιαίτερη προσοχή πρέπει να δοθεί στην επιλογή των οδών στις οποίες γίνεται καταγραφή κυκλοφοριακών δεδομένων. Θα πρέπει να είναι κεντρικές οδικές αρτηρίες, έτσι ώστε να χρησιμοποιούνται από ικανοποιητικό αριθμό χρηστών αλλά και για πολλαπλές διαδρομές.

Σημαντικό συμπέρασμα που προέκυψε και το οποίο εκμεταλλεύεται η εφαρμογή είναι η ύπαρξη μοτίβου στα κυκλοφοριακά δεδομένα, με βάση τη χρονική στιγμή που έχουν αυτά συλλεχθεί. Συγκεκριμένα, η εφαρμογή εκμεταλλεύεται το γεγονός ότι η κυκλοφοριακή συμφόρηση παρουσιάζεται συγκεκριμένες μέρες και ώρες μέσα στην εβδομάδα και αντίστοιχα η κίνηση των οχημάτων διεξάγεται ομαλά κάποιες άλλες ώρες. Για παράδειγμα, τα μεσημέρια των καθημερινών εμφανίζεται αυξημένη κίνηση σε συγκεκριμένες οδικές αρτηρίες. Με τη χρήση της εφαρμογής αυτής, εφόσον έχει γίνει η απαραίτητη συλλογή δεδομένων ο χρήστης μπορεί να αποφύγει τη δυσάρεστη κατάσταση ενός μποτιλιαρίσματος. Πρόκειται ουσιαστικά για έναν συνοδηγό με αρκετή εμπειρία από τις κυκλοφοριακές συνθήκες της πόλης, ώστε να μπορεί να κάνει επιτυχείς προβλέψεις για την κίνηση των δρόμων οποιαδήποτε ώρα της ημέρας.

Όσον αφορά στο τεχνικό τομέα της ανάπτυξης της εφαρμογής καταλήγουμε στα ακόλουθα συμπεράσματα:

Γίνεται εμφανής η σημασία της εύκολης επικοινωνίας μεταξύ διαφορετικών πλατφορμών. Δηλαδή, η εφαρμογή αναπτύχθηκε πάνω σε διαφορετικές πλατφόρμες με κάθε μία από αυτές να είναι υπεύθυνη για κάποιες ξεχωριστές λειτουργίες. Η εφαρμογή, όπως περιγράφεται παραπάνω, αποτελείται από έναν apache server (με βάση δεδομένων mySQL) και μία android εφαρμογή.

Η επικοινωνία μεταξύ server και εφαρμογής android γίνεται με μηνύματα JSON, επιτυγχάνοντας με αυτόν τον τρόπο το τελικό αποτέλεσμα που είναι η σωστή λειτουργία της εφαρμογής. Φαίνεται, έτσι, η αξία της ανάπτυξης προτύπων και πρωτοκόλλων, ούτως ώστε εύκολα να επιτυγχάνεται η επικοινωνία μεταξύ εφαρμογών που πιθανώς έχουν αναπτυχθεί σε διαφορετικά περιβάλλοντα με διαφορετικές γλώσσες προγραμματισμού. Αυτό, εκτός από τη βελτίωση της αποδοτικότητας της συνολικής εφαρμογής, συμβάλει και στην επεκτασιμότητα της, καθώς μπορεί εύκολα να προστεθούν επιπλέον απομακρυσμένες λειτουργίες.

Σημαντικό στοιχείο της εφαρμογής είναι η χρήση του Google Maps API, το οποίο προσφέρει βασικές λειτουργίες χαρτών. Γενικά, μια διεπαφή προγραμματισμού εφαρμογών (application programming interface-API) καθορίζει πώς ορισμένα στοιχεία λογισμικού θα πρέπει να αλληλεπιδρούν μεταξύ τους. Στην πράξη, στις περισσότερες των περιπτώσεων ένα API είναι μια βιβλιοθήκη που συνήθως περιλαμβάνει προδιαγραφές για τη χρήση ρουτινών, δομών δεδομένων, κλάσεων αντικειμένων, και μεταβλητών. Ουσιαστικά πρόκειται για εργαλεία τα οποία αναπτύσσονται με σκοπό την μετέπειτα χρήση του από του ίδιους τους δημιουργούς τους είτε από άλλους. Με αυτόν τον τρόπο ευνοείται η δημιουργία νέων εφαρμογών και τεχνολογιών, επεκτείνονται υλοποιήσεις με αποτέλεσμα τη δημιουργία εφαρμογών που προσφέρουν νέες υπηρεσίες, πηγαίνοντας την επιστήμη της πληροφορικής πολλά βήματα πιο μπροστά.

Για τη σωστή ανάπτυξη της εφαρμογής, αλλά και για την επαλήθευση της σωστής λειτουργίας της εκτελέστηκαν πολλαπλά πειράματα. Σκοπός των πειραμάτων αυτών, πέρα από την επιβεβαίωση των σωστών αποτελεσμάτων της εφαρμογής και την αποσφαλμάτωση της, ήταν η εύρεση των ορίων της εφαρμογής, δηλαδή το μέγεθος των δεδομένων για τα οποία δουλεύει αξιόπιστα. Παράλληλα, μέσω των πειραμάτων, βρέθηκαν οι πραγματικές ανάγκες της εφαρμογής για χρόνο και μνήμη. Η εκτέλεση των πειραμάτων επιπλέον βοήθησε στη διόρθωση σφαλμάτων της εφαρμογής τα οποία δεν είχαν γίνει αντιληπτά πιο πριν. Έτσι, αποδεικνύεται η αναγκαιότητα των πειραμάτων μετά την ολοκλήρωση ανάπτυξης μιας εφαρμογής, τα οποία είναι αναπόσπαστο της κομμάτι της διαδικασίας ανάπτυξης λογισμικού και για αυτό πρέπει να δίνετε ιδιαίτερη έμφαση σε αυτά.

Συνοψίζοντας, μέσα από τη μελέτη, το σχεδιασμό και την ανάπτυξή της εφαρμογής που περιγράφεται στο παρόν έγγραφο, μελετήθηκαν και δοκιμάστηκαν πολλές διαφορετικές σύγχρονες τεχνολογίες, ενώ υπήρξε εμβάθυνση σε τομείς της πληροφορικής.

ΑΝΑΦΟΡΕΣ

[1] Global Positioning System Overview - από το πανεπιστήμιο του Colorado: *http://www.colorado.edu/geography/gcraft/notes/gps/gps_f.html*

[2] The Google Directions API: *https://developers.google.com/maps/documentation/directions/?hl=el*

[3] Γενικές Πληροφορίες για τη λειτουργία της υπηρεσίας Google Maps – από το πανεπιστήμιο του Cornell: *http://blogs.cornell.edu/info2040/2011/09/14/google-maps-its-just-one-big-graph/*.

[4] Dijkstra, E. W. (1959). «A note on two problems in connexion with graphs». Numerische Mathematik 1: 269–271. doi:10.1007/BF01386390.

Μαρία – Άννα Γ. Περιδέλη
sdi0800124@di.uoa.gr

Προηγμένα Ευρετικά Διαχώρισης Πεδίων Τιμών σε Προβλήματα Ικανοποίησης Περιορισμών

Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, Τμήμα Πληροφορικής και Τηλεπικοινωνιών
Πανεπιστημιούπολη, Ιλίσια, 15784, Αθήνα, Ελλάς

Περίληψη

Μελετάμε τη συμβολή της Διαχώρισης των Πεδίων Τιμών σε διάφορα προβλήματα ικανοποίησης περιορισμών. Προκειμένου να μειωθεί ο χρόνος και ο χώρος αναζήτησης, εισάγονται νέα προηγμένα ευρετικά για τη διαχώριση και συνεπώς, τη σμίκρυνση των πεδίων τιμών των μεταβλητών. Μελετάται η απόδοση τους με χρήση του επιλυτή NAXOS σε σχέση με άλλες μεθόδους αναζήτησης. Επίσης, τονίζεται η σημασία της διαχώρισης πεδίου τιμών σε ένα συγκεκριμένο πρόβλημα, αυτό της τυχαίας τοποθέτησης ορθογώνιων πλακιδίων, στο οποίο σημειώθηκαν αποδοτικότερα αποτελέσματα.

Λέξεις κλειδιά: Προγραμματισμός με περιορισμούς, Επιλυτής NAXOS, Διαχώριση, πεδίο-α τιμών, Πρόβλημα τυχαίας τοποθέτησης πλακιδίων.

Επιβλέποντες

Παναγιώτης Σταματόπουλος, Καθηγητής
Νικόλαος Ποθητός, Υποψήφιος Διδάκτωρ

1. Εισαγωγή

Έχετε ακούσει για το πλήθος των chips που υπάρχουν πάνω σε μια μητρική πλακέτα; Έχετε βρεθεί μπροστά σε μια γεμάτη αποθήκη και να θέλετε να τοποθετήσετε κάποιο μεγάλο αντικείμενο αλλά να μην χωράει; Όλα αυτά είναι προβλήματα χωροθέτησης, μια υποκατηγορία των Προβλημάτων Ικανοποίησης Περιορισμών της Τεχνητής Νοημοσύνης, τα οποία συναντάμε συχνά μπροστά μας.

Σε αυτή την εργασία, προσπαθούμε να βελτιώσουμε διάφορα επιλύσιμα προβλήματα ικανοποίησης περιορισμών και ειδικά το πρόβλημα της χωροθέτησης σε μια απλή του μορφή. Σε αυτό θα μας βοηθήσει και η ιδέα της διαχώρισης. Συγκεκριμένα, θα δουλέψουμε πάνω στον επιλυτή NAXOS που είναι προγραμματισμένος σε C++ για να επιλύει Προβλήματα Ικανοποίησης Περιορισμών και θα προσπαθήσουμε να επεκτείνουμε το ευρετικό κομμάτι της επιλογής τιμής από ένα πεδίο τιμών μιας μεταβλητής. Για αυτή την επιλογή θα εισάγουμε και θα χρησιμοποιήσουμε την διαχώριση του πεδίου τιμών σε ποικίλου μεγέθους κομμάτια καθώς και την δίκαια διαχώριση πεδίου τιμών μιας μεταβλητής, δηλαδή τον χωρισμό σε πεδία που θα έχουν περισσότερες πιθανότητες να ικανοποιούν τους περιορισμούς στους οποίους μετέχει μια μεταβλητή.

Συγκρίνοντας τις κλασικές μεθόδους αναζήτησης με τις μεθόδους αναζήτησης διαχώρισης πεδίου τιμών μεταβλητών, θα αποδείξουμε ότι η δεύτερη κατηγορία έχει καλύτερη απόδοση σε προβλήματα χωροθέτησης και συγκεκριμένα στο πρόβλημα τυχαίας τοποθέτησης ορθογώνιων πλακιδίων.

2. Προγραμματισμός με Περιορισμούς

Μια σημαντική κατηγορία προβλημάτων με τα οποία ασχολείται η Τεχνητή Νοημοσύνη, και συγκεκριμένα ένα πεδίο της, ο προγραμματισμός με περιορισμούς, είναι τα προβλήματα ικανοποίησης περιορισμών στα οποία η λύση είναι ένα σύνολο από τιμές που ικανοποιούν διάφορους περιορισμούς.

Ο ακριβής ορισμός τους περιλαμβάνεται μέσα στο εξής τρίπτυχο [1]:

- Περιορισμένες μεταβλητές (variables), που αντιπροσωπεύονται με το σύνολο $\boldsymbol{V} = \{V_1, V_2, \ldots, V_n\}$.
- Πεδία τιμών των μεταβλητών (domains), που αντιπροσωπεύονται με το σύνολο

$\boldsymbol{D} = \{D_1, D_2, \ldots, D_n\}$. Το πεδίο τιμών D_i περιέχει όλες τις επιτρεπόμενες τιμές που μπορούν να δοθούν στη μεταβλητή V_i.

- Περιορισμοί μεταξύ των μεταβλητών (constraints), που αντιπροσωπεύονται με το σύνολο $\boldsymbol{C} = \{C_1, C_2, \ldots, C_m\}$. Κάθε C_i περιλαμβάνει μια σχέση μεταξύ των πεδίων τιμών ενός συνόλου μεταβλητών S_i που ανήκουν στο V, οι οποίες συμμετέχουν στον περιορισμό.

Ως λύση ενός προβλήματος ικανοποίησης περιορισμών ορίζεται μια ανάθεση (assignment) σε όλες τις μεταβλητές V_i του προβλήματος (πλήρης ανάθεση), έτσι ώστε να ικανοποιούνται όλοι οι περιορισμοί C_i του προβλήματος, δηλαδή να είναι όλοι αληθείς (συνεπής ανάθεση).

2.1. Ο Επιλυτής NAXOS

Ο NAXOS SOLVER [2] είναι μια προγραμματιστική βιβλιοθήκη υλοποιημένη στην αντικειμενοστραφή γλώσσα C++, η οποία στοχεύει στην επίλυση προβλημάτων ικανοποίησης περιορισμών. Η χρησιμότητα της έγκειται στο γεγονός ότι μπορεί να λύσει οποιοδήποτε πρόβλημα ικανοποίησης περιορισμών αν εκφραστεί με τον κατάλληλο τρόπο, δηλαδή διαχωρίζει τον τρόπο επίλυσης, ο οποίος είναι ίδιος για όλα τα προβλήματα από τα ζητούμενα ενός προβλήματος. Ο επιλυτής δέχεται ως εισόδους τις μεταβλητές του προγράμματος οι οποίες πρέπει να λάβουν τιμή, καθώς και τους περιορισμούς (εκφρασμένοι ως μαθηματικές σχέσεις στην C++) και παράγει σαν έξοδο μια λύση, δηλαδή τις τιμές των μεταβλητών που ικανοποιούν όλους τους περιορισμούς.

3. Αναζήτηση στα Προβλήματα Ικανοποίησης Περιορισμών

Στον τομέα της Τεχνητής Νοημοσύνης έχουν αναπτυχθεί ποικίλες μορφές αναζήτησης σε μια προσπάθεια γρήγορης και βέλτιστης επίλυσης μεγάλων και δύσκολων προβλημάτων, όπως η βέλτιστη κάλυψη σήματος σε ασύρματα δίκτυα τηλεπικοινωνιών. Προτού όμως προταθεί μια λύση για κάθε πρόβλημα είναι απαραίτητη η σωστή διατύπωση του προβλήματος, τα δεδομένα του δηλαδή αλλά και τα ζητούμενα του [1].

- Η αρχική κατάσταση, δηλαδή η αναπαράσταση της αρχικής εικόνας του προβλήματος σε μια κατανοητή γλώσσα προγραμματισμού ή συμβόλων.

- Η συνάρτηση διαδόχων, η οποία παίρνει ως όρισμα μια κατάσταση x του προβλήματος και επιστρέφει ένα σύνολο (ενέργεια, διάδοχος) που προσδιορίζει τις δυνατές επόμενες «διαδοχικές» καταστάσεις αν εφαρμοστούν κάποιες συγκεκριμένες «ενέργειες».
- Ο έλεγχος στόχου που καθορίζει τις καταστάσεις στόχου, δηλαδή ποιες καταστάσεις δεχόμαστε ως τελικές ενός προβλήματος.
- Η συνάρτηση κόστους μονοπατιού, που είναι το άθροισμα του κόστους όλων των ενεργειών μιας μετάβασης από μία κατάσταση σε μία άλλη.

Στα προβλήματα ικανοποίησης περιορισμών έχουμε μετάβαση από μία κατάσταση σε άλλη όταν επιλέγεται μεταβλητή και της ανατίθεται μια τιμή. Τελική κατάσταση (κατάσταση στόχου) είναι αυτή που έχει όλες τις μεταβλητές δεσμευμένες με κάποια τιμή, ενώ ταυτόχρονα ικανοποιούνται οι περιορισμοί του προβλήματος.

Τα προβλήματα ικανοποίησης περιορισμών τα οποία εξετάσαμε σε αυτή την εργασία είναι πολύ κλασικά στην Τεχνητή Νοημοσύνη. Πρόκειται για: α) το πρόβλημα του πολλαπλασιασμού, όπου θέλουμε να πολλαπλασιάσουμε δυο τριψήφιους αριθμούς, έτσι ώστε σε όλη την ανάλυση και τη διαδικασία του πολλαπλασιασμού κάθε ψηφίο να εμφανίζεται ακριβώς δύο φορές, β) το πρόβλημα των κοινωνικών παικτών του γκολφ με σκοπό την τοποθέτηση *g x s* παικτών σε *g* ομάδες, με *s* άτομα η κάθε μία, για μια περίοδο *w* εβδομάδων, έτσι ώστε κάθε παίκτης να βρίσκεται κάθε βδομάδα σε μια ομάδα με διαφορετικούς παίκτες από όλες τις προηγούμενες που συμμετείχε, γ) το πρόβλημα του χρωματισμού ενός γράφου, ώστε κάθε κόμβος να έχει διαφορετικό χρώμα από τον γειτονικό του και να έχουμε τον ελάχιστο αριθμό χρωμάτων, δ) το πρόβλημα του διαμερισμού αριθμών σε υποσύνολα, ώστε τα αθροίσματα των δύο υποσυνόλων να έχουν την ελάχιστη διαφορά, ε) το πρόβλημα των μαγικών τετραγώνων n-τάξης, στα οποία οποθετούνται οι πρώτοι n^2 φυσικοί αριθμοί με τέτοιο τρόπο ώστε το άθροισμα των n αριθμών σε κάθε γραμμή, στήλη και διαγώνιο να είναι ίσο, στ) το πρόβλημα των N-βασιλισσών, όπου θέλουμε να τοποθετήσουμε N βασίλισσες σε μια σκακιέρα N x N ώστε να μην «συγκρούονται» μεταξύ τους σύμφωνα με τους κανόνες στο σκάκι, ζ) το πρόβλημα της κατασκευής του γνωστού παιχνιδιού Sudoku, όπου έχουμε n^2 τετράγωνα που σχηματίζουν ένα μεγάλο τετράγωνο και σε κάθε μικρό τετράγωνο οι πρώτοι φυσικοί αριθμοί ως n^2 εμφανίζονται ακριβώς μια φορά, όπως και σε κάθε γραμμή και σε κάθε στήλη του μεγάλου τετραγώνου, η) το πρόβλημα της Ελάχιστης Κομβικής Επικάλυψης, που δεδομένου ενός γράφου, ζητείται το ελάχιστο σύνολο των κόμβων ώστε να καλύπτουν κάθε ακμή του γράφου (δηλαδή κάθε ακμή του γράφου να «ακουμπάει» τουλάχιστον έναν

από αυτούς τους κόμβους) και τέλος θ) το πρόβλημα της τυχαίας τοποθέτησης ορθογώνιων πλακιδίων, όπου δημιουργούνται τυχαία μικρά ορθογώνια πλακίδια και πρέπει να τοποθετηθούν σε ένα μεγάλο ορθογώνιο χωρίς να υπερκαλύπτονται μεταξύ τους.

Κάποιες κλασικές μέθοδοι αναζήτησης είναι οι: Πρώτα Κατά Βάθος (DFS), με Περιορισμένη Ασυμφωνία (LDS), Διαμοιρασμού Πίστωσης (credit), Φραγμένου Βάθους με Οπισθοδρόμηση (DBS), με Περιορισμένες Αναθέσεις (LAN), με Φραγμένη Κατά Βάθος Ασυμφωνία (DBDS), με Επαναληπτική Διεύρυνση (Ibroad), με Φραγμένη Οπισθοδρόμηση (BBS), Πρώτα Κατά Βάθος με Επανεκκινήσεις (Rdfs), Ένα-Δείγμα (onesamp), Επαναληπτική Δειγματοληψία (IsampStepping), με Σταδιακό Περιορισμό Πλάτους (GNS) και με Συναρτησιακό Περιορισμό Πλάτους (FNS). Τις δυο τελευταίες αναζητήσεις εισήγαγε ο Φοίβος Θεοχάρης στην πτυχιακή του εργασία [3].

4. Δυο Νέες Μέθοδοι Αναζήτησης Σχετιζόμενες με την Διαχώριση Πεδίου Τιμών Μεταβλητών

4.1. Η Απλή Διαχώριση Πεδίου Τιμών Μεταβλητών

Σε αυτή την εργασία εισάγουμε μια καινούρια ιδέα στην αναζήτηση λύσης στον προγραμματισμό με περιορισμούς, το ευρετικό της Διαχώρισης Πεδίου Τιμών (DomainSplitting).

Ο προβληματισμός που οδήγησε στη συγκεκριμένη ιδέα ήταν το «βάρος» των πιθανών τιμών των μεταβλητών ενός προβλήματος. Σε κάποια προβλήματα, ενώ μια μεταβλητή μπορεί να έχει μεγάλο πεδίο τιμών, από τις λύσεις του προβλήματος παρατηρείται ότι οι επιτρεπτές τιμές της συγκεντρώνονται σε μια περιοχή του πεδίου τιμών, π.χ. σε αυτές που βρίσκονται κοντά στο 0. Ο λόγος που συμβαίνει αυτό είναι η επιβολή περιορισμών στο πρόβλημα, δηλαδή οι περιορισμοί «στενεύουν» τα όρια μιας μεταβλητής που συμμετέχει σε αυτούς. Σε όσους περισσότερους περιορισμούς συναντάται μια μεταβλητή, τόσο πιο πολύ οι τιμές της θα συγκεντρώνονται σε μια μικρή περιοχή του πεδίου τιμών της. Θα μπορούσαμε, λοιπόν, να βελτιώσουμε την απόδοση ενός προβλήματος αν εστιάσουμε την προσοχή μας στην ανάθεση τιμών στις μεταβλητές από μία συγκεκριμένη περιοχή του πεδίου τιμών της ανάλογα με τα ζητούμενα του προβλήματος; Ας εξετάσουμε παρακάτω τον τρόπο λειτουργίας της Διαχώρισης Πεδίου Τιμών.

Το αντικείμενο αυτής της αναζήτησης εστιάζεται στο ευρετικό επιλογής τιμής σε μια περιορισμένη μεταβλητή. Αντί κατά τη διαδικασία της αναζήτησης να αναθέτουμε τιμή σε μια μεταβλητή και να επιβάλλουμε εκ των υστέρων συνέπεια ακμών ώσπου να βρούμε λύση, αποκόπτουμε ένα κομμάτι από το πεδίο τιμών της και κρατάμε τις υπόλοιπες. Έτσι, η ανάθεση δεν γίνεται απ' ευθείας, παρά μόνο όταν έχουν μείνει δύο τιμές στο πεδίο τιμών της μεταβλητής.

Η πιο κοινή διαχώριση πεδίου τιμών είναι στα δύο, δηλαδή η «διχοτόμηση» πεδίου τιμών. Ο λόγος που αποτελεί τη default τιμή για αυτή την αναζήτηση είναι ότι όταν δεν είμαστε σίγουροι ακριβώς για την περιοχή που μαζεύονται οι τιμές των μεταβλητών, είναι πιο εύκολο να προσδιορίσουμε αν είναι στο πρώτο μισό ή στο δεύτερο, ώστε να «καλύπτουμε» πολλές τιμές μαζί. Όταν όμως μπορούμε να προβλέψουμε από πιο πριν τα όρια της περιοχής με τις μαζεμένες επιτρεπτές τιμές για τις μεταβλητές, τότε μπορούμε να διαχωρίσουμε, δηλαδή να «σπάσουμε» τα πεδία τιμών σε ποσοστά, π.χ. στο 10% των πεδίων τιμών αν ξέρουμε ότι οι μικρότερες τιμές είναι πιο πιθανές να οδηγήσουν σε λύση. Αυτή ακριβώς την ιδέα εισάγαμε στη διαχώριση πεδίου τιμών, δηλαδή τη μεταβλητότητα του ποσοστού στο οποίο θα διαχωρίσουμε ένα πεδίο τιμών. Οδηγήσαμε, δηλαδή, την απλή «διχοτόμηση» στη μεταβλητή «διαχώριση».

Η υλοποίηση της διαχώρισης τιμών πεδίων μεταβλητών για προβλήματα ικανοποίησης περιορισμών έγινε ως επέκταση στη βιβλιοθήκη του επιλυτή NAXOS και είχε σαν αποτέλεσμα μια ακόμα μέθοδο αναζήτησης για τον επιλυτή, την goalDomainSplittingLabeling.

4.2. Η «Δίκαια» Διαχώριση Πεδίου Τιμών Μεταβλητών

Η ιδέα για τη διαχώριση του πεδίου τιμών ξεκίνησε με τον προβληματισμό αν οι τιμές των μεταβλητών οι οποίες οδηγούν σε λύση μαζεύονται σε κάποια περιοχή των πεδίου τιμών τους. Ωστόσο, με βάση αυτό δημιουργούνται κάποια νέα ερωτήματα. Είναι σε κάθε μεταβλητή οι «καλές» τιμές της μαζεμένες σε παρόμοιο σημείο με τις άλλες π.χ. στο 10% του πεδίου τιμών της; Και αν με τον όρο «καλές» τιμές εννοούμε αυτές που ικανοποιούν τους περιορισμούς, τότε δεν εξαρτάται από πόσους και ποιους περιορισμούς συμμετέχει η κάθε μεταβλητή; Δεν θα έπρεπε η διαχώριση να έχει σαν αποτέλεσμα δύο πεδία τιμών, που να είναι εξίσου πιθανό να βρίσκεται η λύση ή στο ένα ή στο άλλο και να εξετάζεται πρώτα αυτό με τις λιγότερες τιμές ή μόνο το ένα από τα δύο αν υπάρχει συμμετρία;

Αυτά και κάποια άλλα παρόμοια ερωτήματα έρχονται να μας φέρουν την πρόκληση μιας διαφορετικής διαχώρισης τιμών, πιο «δίκαιας» για τις μεταβλητές ενός προβλήματος ικανοποίησης περιορισμών. Θα μπορούσαμε να ενσωματώσουμε στη διαχώριση πεδίου τιμών ένα διαφορετικό ευρετικό επιλογής τιμής, το οποίο όπως όλα, έχει σαν σκοπό να προτείνει την τιμή από το πεδίο τιμών μιας μεταβλητής που θα οδηγήσει με μεγαλύτερη πιθανότητα σε λύση (succeed-first heuristic) [4].

Υποθέτοντας ότι στα αριθμητικά προβλήματα περιορισμών, αυτό που παίζει σημαντικό ρόλο είναι να είμαστε συνεπείς στα όρια των πεδίων τιμών, αφού τα διαστήματα είναι συνεχή [5], μπορούμε να εφαρμόσουμε ένα διαδεδομένο ευρετικό, αυτό της επιλογής της τιμής για την οποία έχουμε περισσότερες τιμές υποστήριξης (supporters), αλλά σε σύνολα. Αντί δηλαδή να κόβουμε ένα πεδίο τιμών τυχαία σε κάποιο σημείο, θα ήταν ενδιαφέρον να δοκιμάσουμε να το χωρίσουμε σε δυο «ισοβαρή» μέρη όσον αφορά τον αριθμό των τιμών υποστήριξης για τα δυο σύνολα τιμών που θα προκύψουν. Έτσι, θα έχουμε σαν αποτέλεσμα μια πιο «δίκαια» διαχώριση για τις τιμές κάθε μεταβλητής. Δοκιμάσαμε να εξετάσουμε περαιτέρω αυτό το ευρετικό της «δίκαιας» διαχώρισης που σκεφτήκαμε. Να σημειωθεί ότι τα πεδία των τιμών που θα ασχοληθούμε απαρτίζονται από διαδοχικούς ακεραίους.

Συγκεκριμένα, λέμε ότι μια τιμή μεταβλητής έχει «βάρος» ίσο με το πλήθος των τιμών των άλλων μεταβλητών οι οποίες ικανοποιούν τη συγκεκριμένη τιμή στους περιορισμούς στους οποίους εμπλέκεται. Για παράδειγμα, στον περιορισμό $X = Y$, αν το πεδίο τιμών της μεταβλητής X είναι $D_x = \{4, 5, 6\}$ και το πεδίο τιμών της μεταβλητής Y είναι $D_y = \{2, 3, 4\}$, τότε η τιμή 4 στο πεδίο D_x έχει μία τιμή υποστήριξης από το πεδίο D_y (την 4 του D_y), ενώ οι τιμές 5 και 6 δεν έχουν καμία τιμή υποστήριξης. Έτσι, στο ευρετικό θα επιλέγαμε αυτή την τιμή.

Παρατηρούμε, ότι το συγκεκριμένο ευρετικό εξαρτάται από τους περιορισμούς στους οποίους συμμετέχει κάθε μεταβλητή προκειμένου να συγκεντρώσει το πλήθος των «υποστηρικτών» κάθε τιμής του πεδίου τιμών της, ώστε να μπορέσει μετά να κάνει τη διαχώριση στο πεδίο σε ένα ποσοστό που να χωρίζει ισοβαρώς το διάστημα. Από κάθε περιορισμό προκύπτουν ξεχωριστοί κανόνες εύρεσης τιμών υποστήριξης.

Στο πρόβλημα τυχαίας τοποθέτησης ορθογώνιων πλακιδίων πάνω στο οποίο σχεδιάσαμε τη «δίκαια» διαχώριση, η «δικαιοσύνη» έγκειται στο διαφορετικό βάρος κάποιων τιμών των μεταβλητών του προβλήματος, ώστε να εξεταστούν πρώτα οι τιμές που είναι πιο πιθανές να ικανοποιούν τους περιορισμούς. Το

πώς θα διαμορφώσουμε το βάρος μπορούμε να το επηρεάσουμε εκτός από τις παραμέτρους του προβλήματος και από έναν παράγοντα Α, ο οποίος είναι default 50%, και καθορίζει πόσο θέλουμε τα δυο διαστήματα που θα προκύψουν να είναι ισοβαρή στις τιμές υποστήριξης των μεταβλητών τους.

5. Πειραματικά Αποτελέσματα των Μεθόδων Αναζήτησης στα Προβλήματα Ικανοποίησης Περιορισμών

5.1. Σύγκριση της Απλής Διαχώρισης Πεδίου Τιμών Μεταβλητών με τις Κλασικές Μεθόδους Αναζήτησης

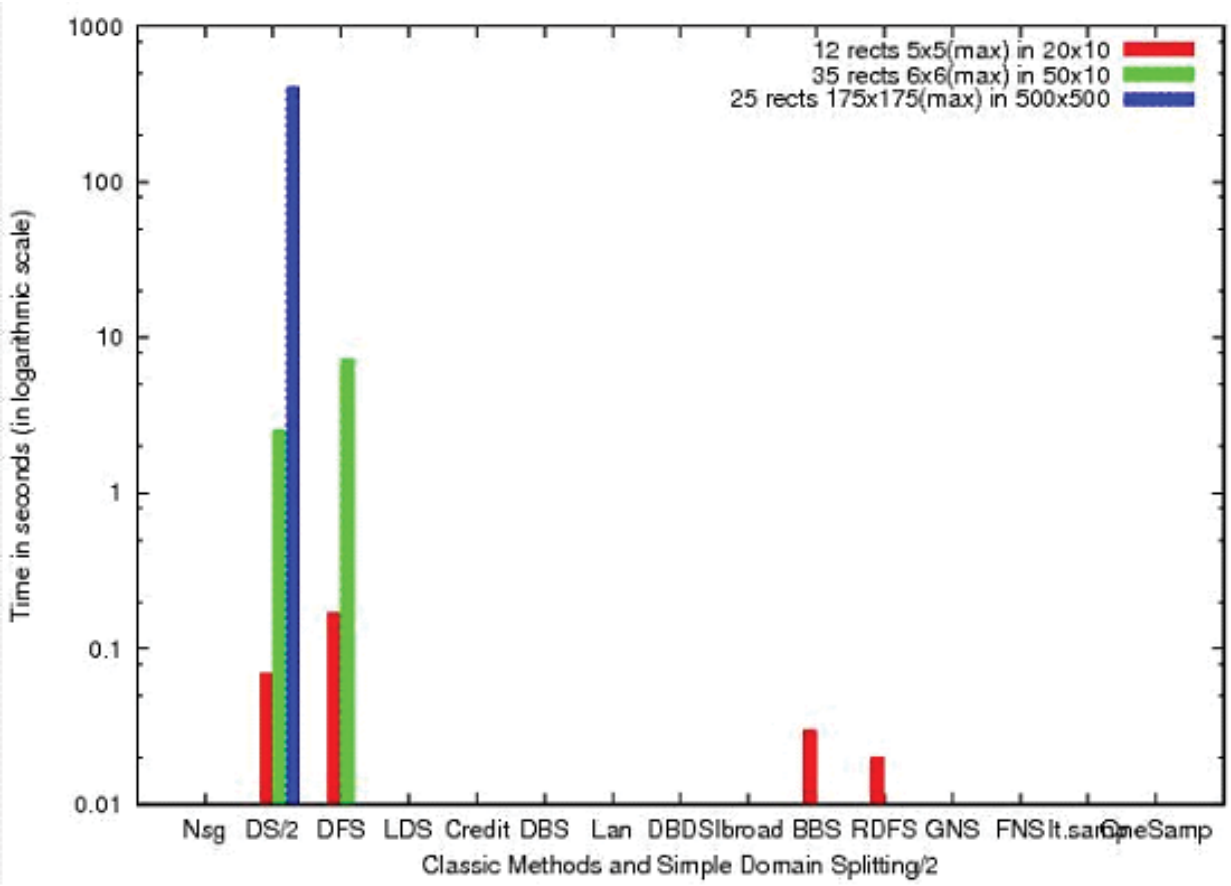
Εξετάσθηκαν διάφορα Προβλήματα Ικανοποίησης Περιορισμών προκειμένου να εντοπιστούν εκείνες οι υποκατηγορίες στις οποίες η Διαχώριση Πεδίου Τιμών θα έχει καλύτερη απόδοση από τις κλασικές μεθόδους αναζήτησης. Η συγκεκριμένη κατηγορία είναι τα προβλήματα χωροθέτησης κι εδώ θα παρουσιάσουμε το πρόβλημα τυχαίας τοποθέτησης ορθογώνιων πλακιδίων (rectangle packing).

Στα υπόλοιπα προβλήματα η διαχώριση πεδίου τιμών συμπεριφέρθηκε ανάλογα με τις κλασικές μεθόδους αναζήτησης, ωστόσο δεν μπόρεσε να ξεπεράσει σε απόδοση 2-3 από αυτές. Πρέπει όμως να τονίσουμε ότι σημείωσε πολύ καλύτερους χρόνους από κάποιες άλλες μεθόδους κι έτσι αξίζει μια θέση ως μέθοδος αναζήτησης στα προβλήματα ικανοποίησης περιορισμών.

Στο πρόβλημα της τυχαίας τοποθέτησης ορθογώνιων πλακιδίων, όπως αναφέραμε και πιο πάνω, το ακριβές ζητούμενο είναι η τοποθέτηση ενός συνόλου τυχαίων παραγόμενων ορθογωνίων διάφορων μεγεθών σε ένα μεγαλύτερο ορθογώνιο (περιοχή τοποθέτησης) με τρόπο ώστε τα ορθογώνια να μη «συγκρούονται» μεταξύ τους και όλες οι πλευρές των ορθογωνίων να είναι παράλληλες προς τις πλευρές της περιοχής τοποθέτησης [6]. Τα δεδομένα του προγράμματος που υλοποιεί το πρόβλημα είναι οι διαστάσεις της περιοχής τοποθέτησης, ο επιθυμητός αριθμός των ορθογωνίων που θέλουμε να τοποθετήσουμε μαζί με ένα άνω όριο στις διαστάσεις τους και η μέθοδος που θα εφαρμοστεί στην αναζήτηση του προβλήματος (στην περίπτωση της διαχώρισης πεδίου τιμών εισάγεται και το ποσοστό της διαχώρισης).

Τα αποτελέσματα της σύγκρισης των κλασικών μεθόδων αναζήτησης και της απλής διαχώρισης πεδίου τιμών μεταβλητών (domain splitting – DS) με το default ποσοστό της διαχώρισης 50% (DS/2) για 3 στιγμιότυπα του προβλήματος είναι τα

εξής:



Σχήμα 1: Ο χρόνος εκτέλεσης του προγράμματος των κλασικών μεθόδων και της απλής διαχώρισης σε δευτερόλεπτα για 3 διαφορετικά στιγμιότυπα, με την τιμή 0 στις μεθόδους να θεωρείται αποτυχία.

Να σημειωθεί εδώ πως ο άξονας των y είναι σε λογαριθμική κλίμακα ώστε να φαίνονται όλες οι τιμές λόγω του μεγάλου εύρους τους στην κλίμακα του χρόνου. Επίσης, για την τρίτη μέτρηση όπου οι άλλες μέθοδοι αποτυγχάνουν, η «διχοτόμηση» πεδίου τιμών κάνει 404.8 δευτερόλεπτα.

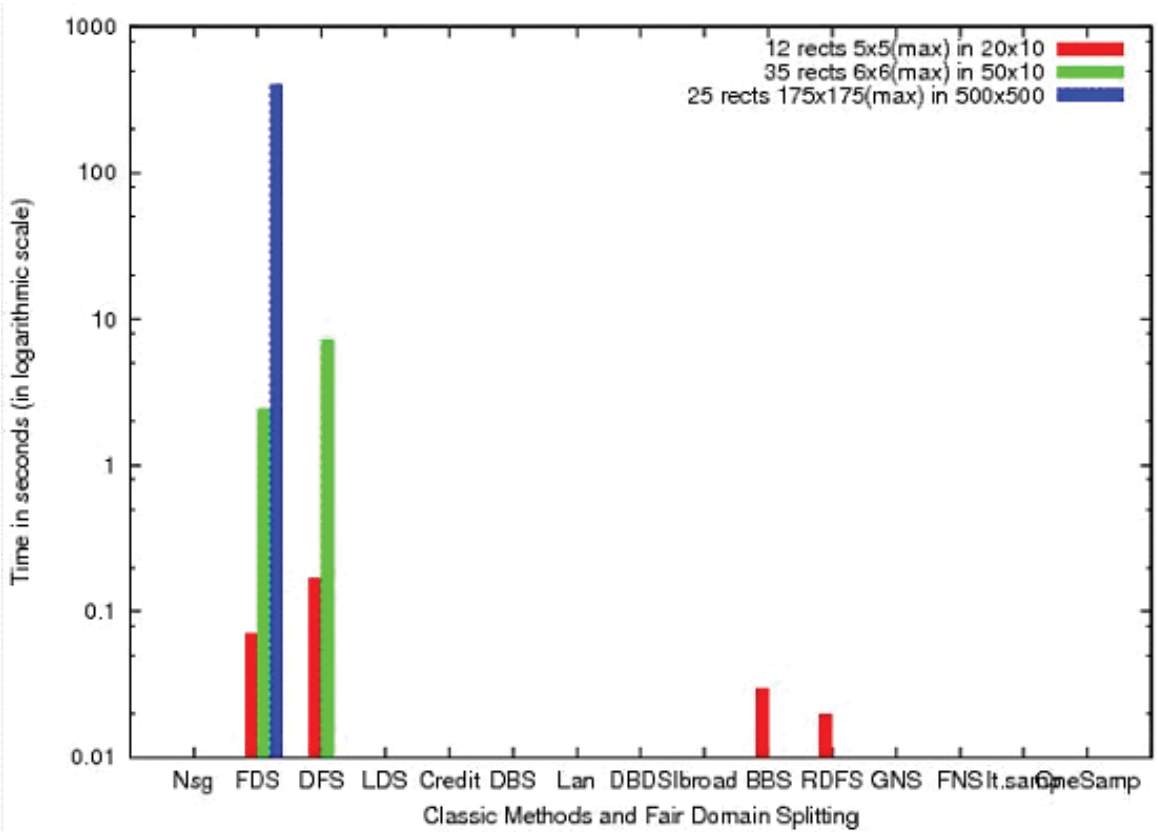
Μετά την εισαγωγή της διαχώρισης πεδίου τιμών, φαίνεται η διαφορά στο πρόβλημα τυχαίας τοποθέτησης, καθώς λύνονται δύσκολα στιγμιότυπα που οι κλασικές μέθοδοι αναζήτησης δεν μπορούσαν να λύσουν σε λογικό χρόνο. Είναι εμφανής η διαφορά στο χρόνο εκτέλεσης στο μεσαίο στιγμιότυπο σε σχέση με την μέθοδο αναζήτησης DFS.

5.2. Σύγκριση της «Δίκαιας» Διαχώρισης Πεδίου Τιμών Μεταβλητών με τις Κλασικές Μεθόδους Αναζήτησης

Προτού προχωρήσουμε στην παράθεση των αποτελεσμάτων, πρέπει να τονίσουμε ότι δεν πρόκειται για μια πλήρη σύγκριση των μεθόδων, γιατί η «δίκαια» διαχώριση διαμορφώνεται διαφορετικά σε κάθε πρόβλημα ανάλογα με τις μεταβλητές και τους περιορισμούς που υπάρχουν, ενώ η απλή διαχώριση και οι κλασικές μέθοδοι εφαρμόζονται σε κάθε πρόβλημα μιας και είναι ανεξάρτητες από τους περιορισμούς του προβλήματος.

Μιας και η διαχώριση πεδίου τιμών φάνηκε να έχει καλά αποτελέσματα σε χωροταξικά προβλήματα, όπως το πρόβλημα τυχαίας τοποθέτησης ορθογώνιων πλακιδίων, οι δοκιμές μας περιορίστηκαν στο συγκεκριμένο πρόβλημα, όπου εξετάσαμε τους περιορισμούς του προβλήματος προκειμένου να διατυπώσουμε μια πιο «δίκαιη» τιμή διαχώρισης για κάθε μεταβλητή.

Τα αποτελέσματα της «δίκαιας» διαχώρισης (fair domain splitting – FDS) με το default ποσοστό-Α 50% σε σύγκριση με τις κλασικές μεθόδους για 3 στιγμιότυπα του προβλήματος είναι τα εξής:



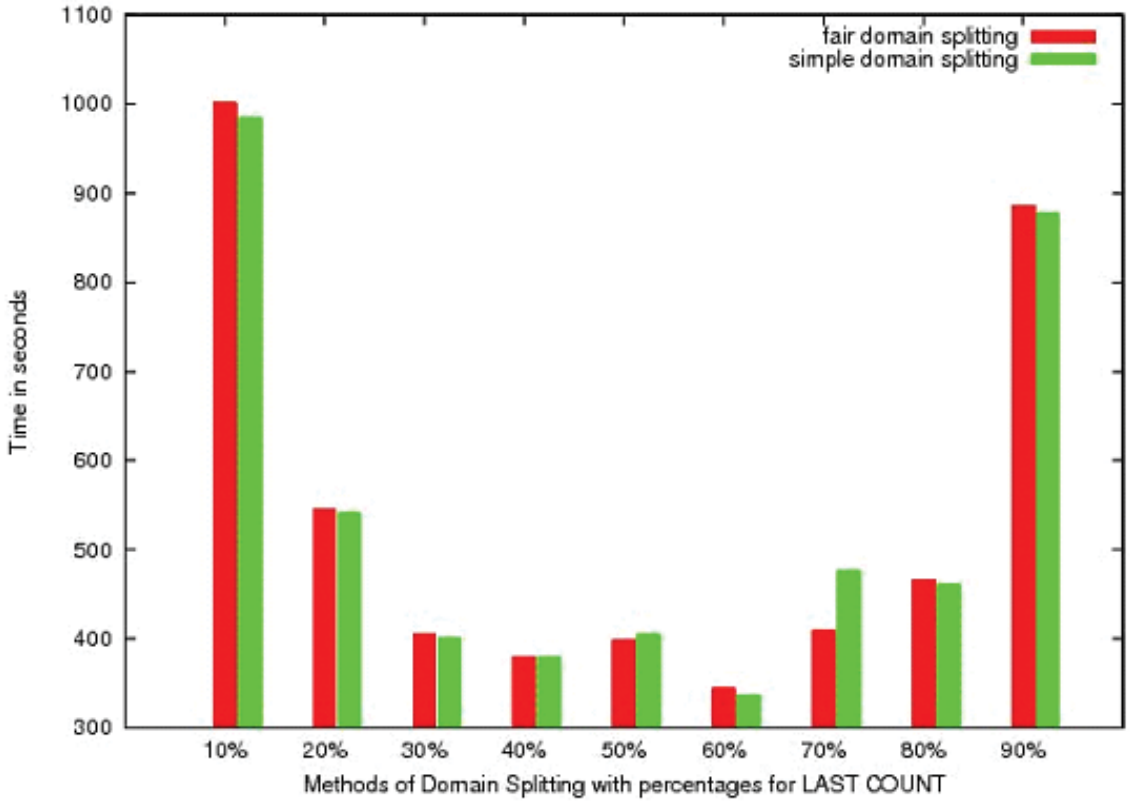
Σχήμα 2: Ο χρόνος εκτέλεσης του προγράμματος των κλασικών μεθόδων και της «δίκαιας» διαχώρισης σε δευτερόλεπτα για 3 διαφορετικά στιγμιότυπα, με την τιμή 0 στις μεθόδους να θεωρείται αποτυχία.

Για μια ακόμα φορά φαίνεται ξεκάθαρα η συμβολή της διαχώρισης πεδίου τιμών, και σε αυτή την εκδοχή της, στο πρόβλημα της τυχαίας τοποθέτησης ορθογώνιων πλακιδίων. Δεν τίθεται, δηλαδή, θέμα ότι η διαχώριση πεδίου τιμών είτε πρόκειται για την απλή είτε για τη «δίκαια» δίνει λύση σε στιγμιότυπα του προβλήματος όπου οι κλασικές μέθοδοι αποτυγχάνουν.

5.3. Σύγκριση των Δυο Μεθόδων Αναζήτησης Διαχώρισης Πεδίου Τιμών Μεταβλητών στο Πρόβλημα Τοποθέτησης Ορθογώνιων Πλακιδίων

Τέλος παρουσίαζουμε μια σύγκριση μεταξύ των δύο ειδών διαχώρισης πεδίου τιμών, της απλής και της «δίκαιας» στο στιγμιότυπο του προβλήματος με τις μεγαλύτερες παραμέτρους, που πλησιάζει περισσότερο τις πραγματικές διαστάσεις του προβλήματος τυχαίας τοποθέτησης ορθογώνιων πλακιδίων.

Τα αποτελέσματα για τα όλα τα ποσοστά των δύο ειδών διαχώρισης σε αυτό το στιγμιότυπο είναι τα εξής:



Σχήμα 2: Ο χρόνος εκτέλεσης του προγράμματος της απλής και της «δίκαιας» διαχώρισης σε δευτερόλεπτα για το 3ο στιγμιότυπο με όλα τα ποσοστά διαχώρισης.

Φαίνεται ξεκάθαρα η συμβολή της «δίκαιας» διαχώρισης πεδίου τιμών στο πρόβλημα όχι μόνο σε σχέση με τις κλασικές μεθόδους, αλλά και σε σχέση με την απλή διαχώριση πεδίου τιμών, στο default (προκαθορισμένο) ποσοστό 50%. Σημειώνουμε την μεγάλη επιρροή των ακραιών ποσοστών στη «δίκαια» διαχώριση, καθώς για μεσαία ποσοστά για τον καθορισμό του Α φαίνεται μια ελαφρώς καλύτερη απόδοση απ' ό,τι στην απλή, ενώ αντίστροφα όσο εφαρμόζουμε ακραίες τιμές ποσοστών η «δίκαια» διαχώριση έπεται της απλής. Ωστόσο και οι δυο μέθοδοι ανεβάζουν το χρόνο τους στα ακραία ποσοστά και έχουν παρόμοιους χρόνους σε αυτά. Εύκολα εξάγεται το συμπέρασμα πως όσο περισσότερο προσαρμόζουμε τη διαχώριση πεδίου τιμών στο πρόβλημα τυχαίας τοποθέτησης ορθογώνιων πλακιδίων (και επιλέγουμε μη ακραία ποσοστά), τόσο περισσότερο μειώνουμε το χρόνο εκτέλεσης του προγράμματος και αυξάνουμε την επίδοσή του.

6. Συμπεράσματα και Μελλοντικές Κατευθύνσεις

Η συνεισφορά αυτής της εργασίας ήταν η εισαγωγή μιας καινούριας μεθόδου αναζήτησης, της διαχώρισης πεδίου τιμών στον επιλυτή προβλημάτων ικανοποίησης περιορισμών NAXOS SOLVER. Παρουσιάσαμε δυο πλευρές αυτής της αναζήτησης, τη γενική μέθοδο διαχώρισης και τη «δίκαια» διαχώριση πεδίου τιμών που προσαρμόζεται ξεχωριστά σε κάθε πρόβλημα και καταλήξαμε σε ένα γνωστό πρόβλημα χωροθέτησης, το πρόβλημα τυχαίας τοποθέτησης ορθογώνιων πλακιδίων, στο οποίο οι δυο μέθοδοι διαχώρισης πεδιου τιμών είχαν καλύτερη απόδοση ενώ οι κλασικές μέθοδοι αναζήτησης αδυνατούσαν να δώσουν λύση.

Η «δίκαια» διαχώριση λόγω του ότι λαμβάνει υπόψιν τους περιορισμούς ενός προβλήματος, σε γενικές γραμμές, είχε μεγαλύτερες αποδόσεις από την απλή, με το μειονέκτημα όμως της μη καθολικότητας. Ίσως επειδή στα μηχανήματα που έγιναν οι μετρήσεις η διαθέσιμη μνήμη έφτανε τα 2GB η κλιμάκωση να μην ήταν πολύ μεγάλη, όμως ενδεχομένως σε μηχανήματα με μεγαλύτερη μνήμη να φαινόταν περισσότερο η συμβολή της «δίκαιας» διαχώρισης στο πρόβλημα τυχαίας τοποθέτησης ορθογώνιων πλακιδίων. Θα ήταν ενδιαφέρον στο μέλλον να επιχειρήσουμε να εφαρμόσουμε αυτές τις δύο μεθόδους σε διάφορες παραλλαγές του προβλήματος τυχαίας τοποθέτησης που προσομοιάζουν ακόμα περισσότερες καταστάσεις της καθημερινότητας ή σε άλλα προβλήματα που ανήκουν στην κατηγορία της χωροθέτησης με περιορισμούς ή ακόμα και σε

άλλες κατηγορίες προβλημάτων ικανοποίησης περιορισμών, όπου η διαχώριση πεδίου τιμών έχει να προσφέρει αποδοτικότερα αποτελέσματα από τις υπόλοιπες μεθόδους αναζήτησης.

Αναφορές

[1] Russel, S., and Norvig, P. (2003) Artificial Intelligence: A modern approach, Pearson Education Inc. (p 179 – 183)

[2] Ποθητός, Ν. (2012) NAXOS SOLVER: Εγχειρίδιο Χρήσης, *http://di.uoa.gr/~pothitos/naxos*

[3] Θεοχάρης, Φ. (2007) Προηγμένες Μέθοδοι Αναζήτησης για Προβλήματα Ικανοποίησης Περιορισμών, Πτυχιακή Εργασία, Τμήμα Πληροφορικής & Τηλεπικοινωνιών, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών.

[4] Bartak, R. (2003) Constraint-based scheduling: An introduction for newcomers, in: Prep. of the 7th IFAC Workshop on Intelligent Manufacturing Systems.

[5] Jussien, N., and Lhomme, O. (1998) Dynamic Domain Splitting for numeric CSP, in: Proc. Eur'n Conf. on Artificial Intelligence, Brightong.

[6] Rudova, H., and Vermirovsky, K. (2004) Random Placement Problem, *http://www.fi.muni.cz/~hanka/rpp/ref.html*

Εμμανουήλ Γ. Τζαγκαράκης
sdi0600143@di.uoa.gr

Νευρωνικά Δίκτυα και Αλγόριθμοι Εκπαίδευσης για Κατηγοριοποίηση Κειμένου

Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, Τμήμα Πληροφορικής και Τηλεπικοινωνιών
Πανεπιστημιούπολη, Ιλίσια, 15784, Αθήνα, Ελλάς

Περίληψη

Σκοπός της παρούσας εργασίας είναι η σύγκριση διαφόρων αλγορίθμων εκπαίδευσης δικτύων πολυεπίπεδων αισθητήρων (multilayer perceptrons) στην εργασία της κατηγοριοποίησης κειμένων. Ελέγχεται η ταχύτητα και το ποσοστό πρόβλεψης που μπορούν να επιτύχουν. Προτείνεται, ακόμα, ένας νέος αλγόριθμος ομαδοποιημένης μάθησης που ονομάζεται σταθερή διάδοση (constant propagation) και δανείζεται στοιχεία τόσο από την κατάβαση πλαγιάς (gradient descent) όσο και από την ελαστική διάδοση (resilient propagation). Ο αλγόριθμος αυτός παρέχει συγκρίσιμα αποτελέσματα τόσο απέναντι στους υπόλοιπους αλγορίθμους εκπαίδευσης όσο και απέναντι σε άλλες μεθόδους κατηγοριοποίησης.

Λέξεις κλειδιά: Νευρωνικά δίκτυα, πολυεπίπεδα δίκτυα αισθητήρων, κατηγοριοποίηση κειμένου, οπισθοδιάδοση, αυξητική οπισθοδιάδοση, οπισθοδιάδοση με ομαδοποιημένη μάθηση, γρήγορη διάδοση, ελαστική διάδοση, σταθερή διάδοση.

Επιβλέπων

Παναγιώτης Σταματόπουλος, Καθηγητής

1. Εισαγωγή

Τα τελευταία χρόνια λόγω του συνεχώς αυξανόμενου όγκου ψηφιοποιημένης πληροφορίας και κατ' επέκταση την ανάγκη για πρόσβαση σε αυτή με διάφορους τρόπους, τα συστήματα ανάκτησης πληροφορίας (information retrieval - IR) έχουν ένα διακεκριμένο ρόλο στον τομέα των πληροφοριακών συστημάτων. Η κατηγοριοποίηση κειμένου (text classification - TC ή text categorization ή topic spotting) είναι μία από τις διαδικασίες των συστημάτων ανάκτησης πληροφορίας [1]. Η μελέτη αυτών των συστημάτων, λοιπόν, στα πλαίσια ενός κόσμου που τα ψηφιακά μέσα διαδραματίζουν όλο και μεγαλύτερο ρόλο στις επιχειρήσεις, αλλά και στην καθημερινότητα, κρίνεται επιτακτική.

Ένας ορισμός της κατηγοριοποίησης κειμένου είναι ο εξής: TC είναι η διαδικασία ανάθεσης μιας δυαδικής τιμής σε κάθε ζευγάρι $(d_j, c_i) \in D \times C$ όπου D είναι ένα πεδίο ορισμού που περιέχει κείμενα που πρέπει να κατηγοριοποιηθούν και το C ένα σύνολο το οποίο περιέχει όλες τις πιθανές κατηγορίες. Όταν η τιμή στο ζευγάρι είναι αληθής (true) τότε αυτό σημαίνει ότι το κείμενο ανήκει στην κατηγορία [1].

Στη δεκαετία του 1990 και μετά, προτάθηκαν μέθοδοι κατηγοριοποίησης από μηχανές επιβλεπόμενης και μη μάθησης (supervised/ unsupervised) των οποίων το κύριο χαρακτηριστικό είναι ότι δεν χρειάζονται κάποιον ειδικό για να φτιάξει κανόνες για κατηγοριοποίηση και ότι οι μηχανές αυτές μπορούν να αλλάξουν πεδίο εφαρμογής χωρίς να χρειάζεται να δημιουργηθεί μία νέα δομή κανόνων από ένα ειδικό όπως γινόταν με την προσέγγιση της μηχανικής της γνώσης (knowledge engineering) [2]. Οι κυριότερες μηχανές που έχουν προταθεί στη βιβλιογραφία είναι οι: Πιθανοτικοί κατηγοριοποιητές (Probabilistic), τα δέντρα απόφασης και μάθησης με επαγωγικούς κανόνες (inductive rule learners), οι μέθοδοι παλινδρόμησης (Regression Methods - Linear least Squares fit), η μέθοδος Rocchio, οι κατηγοριοποιητές βασισμένοι σε παράδειγμα (example based Classifiers - kNN) και οι Μηχανές διανύσματος υποστήριξης (Support Vector Machines). Φυσικά σε αυτές τις μεθόδους συμπεριλαμβάνονται και τα Νευρωνικά δίκτυα που είναι η μέθοδος που εστιάζει η παρούσα μελέτη [1][3].

Η παρούσα μελέτη εστιάζει επίσης και στους αλγόριθμους εκπαίδευσης των πολυεπίπεδων νευρωνικών δικτύων. Συγκεκριμένα, μελετώνται οι αλγόριθμοι της ομαδοποιημένης οπισθοδιάδοσης, της οπισθοδιάδοσης με αυξητική μάθηση, της γρήγορης διάδοσης και της ελαστικής διάδοσης [5][6]. Ταυτόχρονα, προτείνεται ένας νέος αλγόριθμος εκπαίδευσης, με ονομασία σταθερή διάδοση,

ο οποίος συνδυάζοντας χαρακτηριστικά από τις παραπάνω μεθόδους προσφέρει υποσχόμενα αποτελέσματα.

2. Νευρωνικά δίκτυα: Δομή, Στοιχεία και Αλγόριθμοι Εκπαίδευσης

Τα νευρωνικά δίκτυα αποτελούνται από απλές υπολογιστικές μονάδες, τους νευρώνες ή κόμβους και τις συνδέσεις μεταξύ τους που ονομάζονται συνάψεις ή βάρη. Αναλυτικότερα τα στοιχεία που αποτελούν ένα νευρωνικό δίκτυο είναι:

- Οι μονάδες επεξεργασίας ή κόμβοι ή νευρώνες.
- Μία ποσότητα ενεργοποίησης του εκάστοτε κόμβου που το αποτέλεσμά της αποτελεί ταυτόχρονα και την έξοδο του κάθε νευρώνα.
- Οι συνδέσεις μεταξύ των κόμβων, γνωστές και ως βάρη ή συνάψεις, που καθορίζουν την ένταση που έχει το σήμα εξόδου από τον κόμβο αναχώρησης προς τον κόμβο προορισμού.
- Ένας κανόνας διάδοσης (Propagation rule), ο οποίος καθορίζει τις εισόδους του κάθε κόμβου και τον τρόπο με τον οποίο διαδίδεται ένα σήμα εισόδου.
- Μία συνάρτηση ενεργοποίησης, η οποία είναι υπεύθυνη να συλλέξει τα δεδομένα από τις εισόδους σε ένα κόμβο και να υπολογίσει την ποσότητα ενεργοποίησης.
- Μια μέθοδος μάθησης [4].

2.1. Δομή Νευρωνικών Δικτύων

Αυτή τη στιγμή υπάρχουν πολλές κατηγορίες και τοπολογίες νευρωνικών δικτύων, ενώ νέες μεθοδολογίες και τοπολογίες ή παραλλαγές παλαιοτέρων ανακαλύπτονται καθημερινά [7][5][16]. Τα νευρωνικά δίκτυα που υπάρχουν στην βιβλιογραφία μπορούν να χωριστούν με διάφορους τρόπους. Ο τρόπος εκπαίδευσης, η τοπολογία και τα δεδομένα που δέχονται είναι κάποιοι από αυτούς. Σχετικά με τον τρόπο εκπαίδευσης, συναντάμε νευρωνικά δίκτυα με επιβλεπόμενο και μη επιβλεπόμενο μηχανισμό μάθησης. Στον επιβλεπόμενο τρόπο μάθησης τα δίκτυα γνωρίζουν τα αποτελέσματα που πρέπει να στοχεύσουν,

ενώ στον μη επιβλεπόμενο τρόπο μάθησης δεν γνωρίζουν τα αποτελέσματα και συχνά επιτελούν εργασίες μείωσης των διαστάσεων ή συμπίεσης δεδομένων. Η τοπολογία των δικτύων μπορεί να χωριστεί σε δύο βασικές κατηγορίες, στα δίκτυα με προς τα εμπρός τροφοδότηση (feed forward δίκτυα) που είναι ακυκλικοί γράφοι και στα αναδρομικά ή με προς τα εμπρός τροφοδότηση νευρωνικά δίκτυα, όπου μπορούν να υπάρξουν κύκλοι στο γράφο. Τέλος, ανάλογα με τα δεδομένα που διαχειρίζονται, έχουμε το διαχωρισμό των νευρωνικών δικτύων σε δίκτυα που είναι κατάλληλα για κατηγοριοποίηση, όπου υπάρχουν μεταβλητές κατηγοριών που μπορούν να πάρουν πεπερασμένες τιμές και στην κάθε κατηγορία ανήκουν πολλά παραδείγματα τα οποία μπορούμε να τα βρούμε τόσο με επιβλεπόμενους όσο και μη επιβλεπόμενους μηχανισμούς μάθησης, και σε δίκτυα όπου το αποτέλεσμα τους μπορεί να είναι ένας αριθμός που αντιπροσωπεύει τη μέτρηση ενός μεγέθους, όπως πχ το μήκος ενός αντικειμένου. Η διαδικασία επιβλεπόμενης μάθησης σε τέτοια δίκτυα ονομάζεται παλινδρόμηση [5][8][9]. Η παρούσα μελέτη θα επικεντρωθεί στη δομή και τις τεχνικές μάθησης των δικτύων πολυεπίπεδων αισθητήρων σταθερής τοπολογίας που είναι προς τα εμπρός τροφοδοτούμενα δίκτυα επιβλεπόμενης μάθησης.

2.2. Εκπαίδευση των πολυεπίπεδων νευρωνικών δικτύων

Ο κανόνας μάθησης αισθητήρων χρησιμοποιεί τη διαφορά μεταξύ του επιθυμητού αποτελέσματος και του παραγόμενου από το δίκτυο αισθητήρων αποτελέσματος για να αλλάξει τα βάρη σύμφωνα με τον κανόνα. Μία τέτοια προσέγγιση μπορεί να εφαρμοστεί στους κόμβους εξόδου ενός πολυεπίπεδου δικτύου αισθητήρων, αλλά όχι στα εσωτερικά κρυμμένα επίπεδα μιας και για αυτά δεν υπάρχει κάποιο αποτέλεσμα στόχος για να συγκρίνουμε την έξοδό τους. Το παραπάνω πρόβλημα είναι γνωστό και ως πρόβλημα ανάθεσης πίστωσης (credit assignment problem) [5]. Η απάντηση στο παραπάνω πρόβλημα είναι να ορίσουμε μία συνάρτηση που μετράει το λάθος καθολικά, όπως η συνάρτηση του μέσου τετραγωνικού λάθους (Mean Square Error – MSE) η οποία είναι το άθροισμα των τετραγώνων των διαφορών μεταξύ του επιθυμητού αποτελέσματος και αυτού που παράγει το δίκτυο. Η συνάρτηση MSE είναι μία διαφοροποιήσιμη συνάρτηση των εξόδων του δικτύου και η τιμή της, το λάθος δηλαδή, είναι μία διαφοροποιήσιμη συνάρτηση των βαρών του δικτύου. Συνεπώς, μπορούμε με διάφορες μεθόδους να βρούμε τα βάρη που ελαχιστοποιούν τη συνάρτηση του λάθους. Τέτοιες μέθοδοι είναι η κατάβαση πλαγιάς (gradient descent), καθώς και άλλοι πιο αποτελεσματικοί τρόποι βελτιστοποίησης [5]. Η συνάρτηση λάθους

που προσπαθούμε να ελαχιστοποιήσουμε είναι η:

$$E_p = \frac{1}{2}\sum_{i=1}^{n}(o_i - t_i)^2 \qquad (1)$$

όπου E_p είναι το συνολικό λάθος, n είναι ο αριθμός των παραδειγμάτων εκπαίδευσης t_i είναι το αποτέλεσμα που στοχεύουμε και o_i είναι το αποτέλεσμα που παράγει το δίκτυο. Το 0.5 στην αρχή του τύπου χρησιμοποιείται για την απλούστευση της παραγώγου. Αναλυτικότερα θέλουμε να υπολογίσουμε το:

$$\nabla E = \left(\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \ldots, \frac{\partial E}{\partial w_l}\right) \qquad (2)$$

ώστε τελικά να υπολογίσουμε το Δw, την αλλαγή δηλαδή του κάθε βάρους με τον παρακάτω τύπο:

$$\Delta w_i = -\gamma \frac{\partial E}{\partial w_i} \qquad (3)$$

για i από 1 έως l, όπου γ είναι μία σταθερά που λέγεται ρυθμός μάθησης με προτεινόμενες τιμές στο διάστημα [0.25-0.5] [6]. Σκοπός της διαδικασίας τελικά είναι να μηδενίσουμε τη συνάρτηση λάθους.

Σχετικά με τη διαδικασία της μάθησης στα νευρωνικά δίκτυα έχουν προταθεί διάφοροι αλγόριθμοι με του κυριότερους να είναι ο αλγόριθμος οπισθοδιάδοσης (backpropagation - bp) με ομαδοποιημένη (batch) [6] και με αυξητική (online) μάθηση, ο αλγόριθμος γρήγορης διάδοσης (qprop) [10] καθώς και ο αλγόριθμος ελαστικής διάδοσης (resilient propagation - rprop) [11].

2.3. Πρόταση για ένα βελτιωμένο αλγόριθμό εκπαίδευσης νευρωνικών δικτύων: Σταθερή διάδοση.

Η σταθερή διάδοση είναι μία μέθοδος ομαδοποιημένης μάθησης. Ο αλγόριθμος που προτείνεται έχει δανειστεί στοιχεία από την ελαστική διάδοση, τις διάφορες μεθόδους καθολικής προσαρμογής του ρυθμού μάθησης, αλλά και από τις διάφορες βελτιώσεις της οπισθοδιάδοσης. Η ιδέα που προέρχεται από την

ελαστική διάδοση είναι ότι το βήμα που κάνει ο αλγόριθμος για το κάθε βάρος θα πρέπει να είναι ανεξάρτητο από το απόλυτο μέγεθος της κλίσης. Σε αντίθεση με την ελαστική διάδοση δεν χρησιμοποιήθηκε η αλλαγή του πρόσημου της κλίσης μεταξύ της τωρινής και της προηγούμενης εποχής, παρά μόνο το πρόσημο που έχει η κλίση στην παρούσα εποχή. Η ποσότητα της οποίας ελέγχεται το πρόσημο είναι η:

$$\Delta w(t) = -\gamma * \frac{\partial \mathrm{E}}{\partial w} + momentum * \Delta w(t-1) \qquad (4)$$

Ανάλογα με το πρόσημο της παραπάνω ποσότητας, δηλαδή το *sign(Δw(t)),* εφαρμόζεται μία σταθερού μεγέθους αλλαγή πάνω στο εκάστοτε βάρος, έστω Step. Αποφασίστηκε να μην εφαρμοστεί κάποια τεχνική αύξησης ή μείωσης του μεγέθους του Step, ή καλύτερα όχι με τον τρόπο τον οποίο το χρησιμοποιεί η ελαστική διάδοση, και αυτός είναι ένας ακόμη λόγος που δεν χρησιμοποιήθηκε άμεσα το πρόσημο της κλίσης στην προηγούμενη εποχή. Η αιτιολογία για το παραπάνω είναι ότι οι δυναμικές μεταβολές του βήματος που προκαλεί η ελαστική διάδοση επιταχύνουν αρχικά τη διαδικασία, αλλά αποτυγχάνουν να κρατήσουν εγγυημένα ένα σταθερά καλό αποτέλεσμα όταν το καθολικό λάθος έχει μειωθεί αρκετά. Αυτό το χαρακτηριστικό γίνεται προσπάθεια να επιτευχθεί με τη σταθερή διάδοση θυσιάζοντας βέβαια με αυτόν τον τρόπο το ιδανικό βήμα που στη θεωρία υπολογίζει τόσο η ελαστική όσο και η γρήγορη διάδοση. Ο αλγόριθμος έχει ως εξής:

```
Στο τέλος κάθε εποχής για κάθε βάρος και κόμβο πόλωσης {
     bραλλαγή_ij=-γ*(∂E/(∂w_ij))+momentum*Δw(t-1)

     αν (bραλλαγή_ij> 0.0)  τότε:
          Δw_ij (t) = (η_plus)*AT

     αν (bραλλαγή_ij< 0.0)  τότε:
          Δw_ij (t)=[-η_minus)]*AT
}
```

3. Αποτελέσματα

Στην παράγραφο αυτή παρατίθενται τα αποτελέσματα της σύγκρισης των διαφόρων αλγορίθμων για διάφορα σετ δεδομένων μετά από προσεκτική επιλογή διαφόρων παραμέτρων λειτουργίας του κάθε ενός για το εκάστοτε σύνολο.

3.1. Σύνολο δεδομένων κριτικών χρηστών του Imdb

Το σύνολο αποτελείται από 692 αρνητικές και 694 θετικές κριτικές ταινιών. Το σύνολο αυτό μπορεί κάποιος να το βρει στη διεύθυνση:

http://www.cs.cornell.edu/people/pabo/movie-review-data/mix20_rand700_tokens_0211.tar.gz

Ακολουθούν τα αποτελέσματα για 500 εποχές εκπαίδευσης:

500 εποχές	1	2	3	4	5	Μ.Ο.
πρόβλεψη batch bp	0.6892	0.6289	0.6578	0.6699	0.6386	0.6569
πρόβλεψη online bp	0.8289	0.7952	0.8289	0.7904	0.8	0.8087
πρόβλεψη qprop	0.5614	0.5976	0.6241	0.5398	0.5952	0.5836
πρόβλεψη rprop	0.8072	0.7807	0.747	0.7759	0.8	0.7822
πρόβλεψη cprop	0.8289	0.812	0.8193	0.8096	0.8217	0.8183

Πίνακας 1: Αποτελέσματα κριτικών χρηστών του Imdb

3.2. Σύνολο δεδομένων spam και ham μηνυμάτων κειμένου κινητού τηλεφώνου

Το σύνολο αυτό αποτελείται από 5.574 μηνύματα κειμένου εκ των οποίων τα 4.827 είναι ham και τα 747 είναι spam. Μπορεί να βρεθεί εδώ:

http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/.

Εξαιρετικά σημαντικό στην κατηγοριοποίηση spam και ham δεν είναι μόνο η ταχύτητα σύγκλισης και το επίπεδο της ακρίβειας που θα επιτευχθεί, αλλά και το

ποσοστό των μηνυμάτων της κατηγορίας ham που δεν θα κατηγοριοποιηθούν σαν spam, γεγονός που θα έχει σαν αποτέλεσμα στον πραγματικό κόσμο ο χρήστης να μην δει ένα κανονικό μήνυμα. Τα αποτελέσματα θα είναι ο μέσος όρος πέντε προσπαθειών.

200 εποχές	1	2	3	4	5	Μ.Ο.
πρόβλεψη online bp	0.9755	0.9815	0.9833	0.985	0.9803	0.9811
Λάθος εκπαίδευσης	0.0009	0.0006	0.0004	0.0005	0.0007	0.0006
Ποσοστό ham	0.9867	0.9965	0.9923	0.998	0.9931	0.9933
Ποσοστό spam	0.9113	0.887	0.927	0.8957	0.8982	0.9038
πρόβλεψη batch bp	0.9551	0.9258	0.9557	0.9528	0.9659	0.9511
Λάθος εκπαίδευσης	0.0207	0.0133	0.0113	0.0176	0.016	0.0158
Ποσοστό ham	0.9717	0.9938	0.9677	0.9735	-	0.9767
Ποσοστό spam	0.8475	0.4796	0.8767	0.827	-	0.7577
πρόβλεψη qprop	0.9557	0.9492	0.9581	0.9522	0.9617	0.9554
Λάθος εκπαίδευσης	0.0119	0.0138	0.0127	0.0134	0.0103	0.0124
Ποσοστό ham	0.9815	0.9896	0.989	0.9903	0.9898	0.988
Ποσοστό spam	0.7751	0.6889	0.7477	0.704	0.7598	0.7351
πρόβλεψη rprop	0.9743	0.9581	0.9767	0.9593	0.9773	0.9691
Λάθος εκπαίδευσης	0.0057	0.011	0.0048	0.0105	0.0063	0.0077
Ποσοστό ham	0.9924	0.9937	0.9932	0.9843	0.9918	0.9911
Ποσοστό spam	0.8559	0.755	0.8578	0.7794	0.8732	0.8243
πρόβλεψη cprop	0.9351	0.951	0.9563	0.9528	0.9378	0.9466
Λάθος εκπαίδευσης	0.0167	0.0112	0.0087	0.012	0.0171	0.0131
Ποσοστό ham	0.9931	0.9979	0.9944	0.974	0.9707	0.986
Ποσοστό spam	0.5664	0.6535	0.7246	0.8038	0.7406	0.6978

Πίνακας 2: Αποτελέσματα συνόλου spam και ham μηνυμάτων κειμένου κινητών τηλεφώνων

Κοιτάζοντας προσεκτικότερα, τα παραπάνω αποτελέσματα ειδικά για την οπισθοδιάδοση με αυξητική μάθηση είναι εξαιρετικά, όπου σε περίπου 1.450 κανονικά μηνύματα δεν κατηγοριοποιεί σωστά μόνο τα κατά μέσο όρο 10 από αυτά, νούμερο που σε ποσοστό είναι ένας πλήρως αποδεκτός αριθμός. Ακόμα, πρέπει να τονιστεί ότι και οι υπόλοιποι αλγόριθμοι λειτούργησαν αξιοπρεπέστατα, καθιστώντας αυτό το πολυεπίπεδο δίκτυο αισθητήρων μία αρκετά ικανοποιητική μηχανή κατηγοριοποίησης sms μηνυμάτων.

3.3. Σύνολο δεδομένων spam και ham μηνυμάτων ηλεκτρονικού ταχυδρομείου

Το σύνολο δεδομένων μηνυμάτων ηλεκτρονικού ταχυδρομείου είναι ένα πολύ σημαντικό τεστ ειδικά για εφαρμογές του πραγματικού κόσμου. Το παρόν σύνολο περιλαμβάνει 4.327 μηνύματα εκ των όποιων κανονικά (ham) είναι τα 2.949 και spam τα 1.378. Το σύνολο μπορεί να το βρει κάποιος στη διεύθυνση:

http://csmining.org/index.php/spam-email-datasets-.html

Ακολουθεί ο πίνακας αποτελεσμάτων:

200 εποχές	1	2	3	4	5	Μ.Ο.
Πρόβλεψη online bp	0.9611	0.6895	0.6672	0.6834	0.6818	0.7366
Λάθος Εκπαίδευσης	0.1129	0.0576	0.0553	0.0582	0.0571	0.0682
Ποσοστό ham	1	1	1	1	1	1
Ποσοστό spam	0	0	0	0.0072	0	0.0014
Πρόβλεψη batch bp	0.7643	0.8413	0.6888	0.829	0.879	0.8005
Λάθος Εκπαίδευσης	0.0833	0.0757	0.0658	0.0564	0.0701	0.0703
Ποσοστό ham	0.99	0.9728	0.9898	0.8209	0.9513	0.945
Ποσοστό spam	0.2594	0.5212	0.0459	0.8447	0.7362	0.4815
Πρόβλεψη qprop	0.7535	0.7365	0.7535	0.7296	0.7227	0.7391
Λάθος Εκπαίδευσης	0.0598	0.0617	0.0599	0.0584	0.0613	0.0602
Ποσοστό ham	0.8849	0.8409	0.8731	0.8669	0.8625	0.8657
Ποσοστό spam	0.4872	0.5121	0.5059	0.4699	0.45	0.485

200 εποχές	1	2	3	4	5	Μ.Ο.
Πρόβλεψη rprop	0.9145	0.8097	0.7712	0.7334	0.8444	0.8146
Λάθος Εκπαίδευσης	0.023	0.0574	0.0577	0.067	0.0339	0.0478
Ποσοστό ham	0.9543	0.7935	0.8843	0.7506	0.8435	0.8452
Ποσοστό spam	0.825	0.8447	0.5388	0.6966	0.8463	0.7503
Πρόβλεψη cprop	0.8891	0.9237	0.896	0.9106	0.9253	0.9089
Λάθος Εκπαίδευσης	0.0169	0.0205	0.0272	0.0241	0.0223	0.0222
Ποσοστό ham	0.9816	0.9276	0.9561	0.9431	0.9781	0.9573
Ποσοστό spam	0.7016	0.9155	0.7653	0.8425	0.8	0.805

Πίνακας 3: Αποτελέσματα συνόλου spam και ham μηνυμάτων ηλεκτρονικού ταχυδρομείου

3.4. Σύνολο δεδομένων 1 από n, σύνολο ιατρικών κειμένων

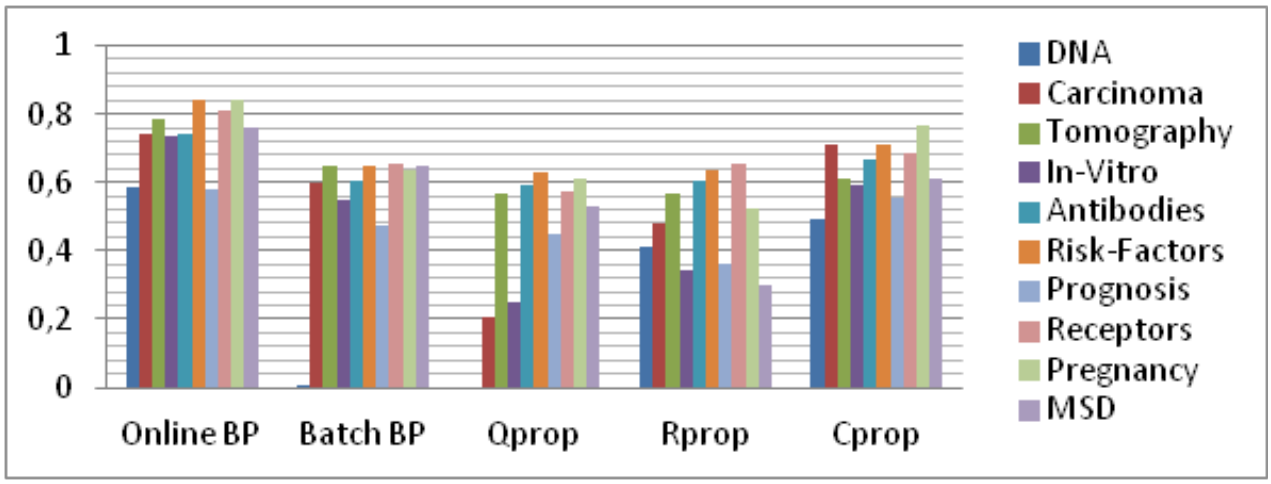
Το παραπάνω συμπιεσμένο αρχείο περιέχει 19 σύνολα δεδομένων. Συγκεκριμένα, 1159 για Antibodies, 709 για Carcinoma, 764 για DNA, 1001 In-Vitro, 864 για Molecular Sequence Data, 1.621 για Pregnancy, 1.037 για Prognosis, 1.297 για Receptors, 1.450 για Risk Factors και 1.260 για Tomography. Το σύνολο μπορεί να βρεθεί στις συλλογές του weka στην ιστοσελίδα:

http://www.cs.waikato.ac.nz/ml/weka/datasets.html.

Συγκεκριμένα είναι το 19MclassTextWc.zip με σύνδεσμο:

http://prdownloads.sourceforge.net/weka/19MclassTextWc.zip?download

Τα αποτελέσματα είναι τα εξής:



Σχήμα 1: Αποτελέσματα κατηγοριοποίησης συνόλου ιατρικών κειμένων

3.5. Συγκριτικά αποτελέσματα με άλλους αλγορίθμους κατηγοριοποίησης

Στο παρόν κεφάλαιο γίνεται μελέτη για το πως συμπεριφέρονται οι καλύτερες εκδοχές ενός πολυεπίπεδου δικτύου αισθητήρων απέναντι στους διαδεδομένους αλγορίθμους κατηγοριοποίησης όπως τα αφελή Bayes δίκτυα και ο SVM. Αναλυτικά:

Για το σύνολο κριτικών έχουμε:

Σύνολο IMDB	Ποσοστό Πρόβλεψης
Σταθερή διάδοση Μ.Ο.	0.8183
Αφελή Bayes Δίκτυα	0.826923
SVM	0.846154

Πίνακας 4: Αποτελέσματα κατηγοριοποίησης για το σύνολο κριτικών ταινιών από το Imdb

Για το σύνολο τον μηνυμάτων κινητού τηλεφώνου και του συνόλου μηνυμάτων ηλεκτρονικού ταχυδρομείου έχουμε:

Σύνολο SMS	HAM SMS	SPAM SMS	Ποσοστό HAM	Ποσοστό SPAM
Σταθερή διάδοση Μ.Ο	0.986	0.6978	0.9515	0.9767
Online BP	0.993	0.9038	-	-
Αφελή Bayes Δίκτυα	0.992	0.913	0.9599	0.978
SVM	0.984	0.981	0.9784	0.983

Πίνακας 5: Αποτελέσματα κατηγοριοποίησης για το σύνολο των μηνυμάτων κειμένου sms και email

Για το σύνολο των ιατρικών κειμένων έχουμε:

Ιατρικό σύνολο	Ποσοστό Πρόβλεψης
Σταθερή διάδοση Μ.Ο.	0.718937
Αφελή Bayes Δίκτυα	0.733055
SVM	0.763511
online BP	0.780167

Πίνακας 6: Αποτελέσματα κατηγοριοποίησης για το σύνολο των ιατρικών κειμένων

4. Συμπεράσματα

Μελετώντας προσεκτικά τη συμπεριφορά και την απόδοση των πολυεπίπεδων δικτύων αισθητήρων με διάφορους αλγορίθμους και τοπολογίες προκύπτουν ενδιαφέροντα συμπεράσματα. Αρχικά, όσον αφορά την απόδοση των πολυεπίπεδων δικτύων αισθητήρων μπορούμε να πούμε ότι η εκπαίδευσή τους είναι αργή συγκριτικά με αυτή των ανταγωνιστικών αλγορίθμων. Το γεγονός αυτό κάνει επιτακτική την ανάγκη μελέτης οποιουδήποτε τρόπου βελτίωσης της ταχύτητας με την οποία μαθαίνουν, ώστε το δίκτυο να μπορεί να εκπαιδεύεται

γρηγορότερα στα συνήθως πολλών διαστάσεων διανύσματα εισόδου που χρησιμοποιούνται στην κατηγοριοποίηση κειμένων. Τέτοιες μέθοδοι βελτιστοποίησης μπορεί να αφορούν αυστηρά την υλοποίηση, με τις οποίες δεν ασχολήθηκε η παρούσα εργασία, όσο και με μεθόδους που θα κάνουν το εκάστοτε βήμα του αλγορίθμου πιο αποδοτικό, επιταχύνοντας την πορεία προς το ελάχιστο της συνάρτησης λάθους. Τα αποτελέσματα δείχνουν ξεκάθαρα ότι ο πιο απλός τρόπος για να επιτευχθεί αυτό είναι η χρησιμοποίηση δικτύου ενός κρυφού επιπέδου με κατάλληλο αριθμό κόμβων, χρησιμοποιώντας οπισθοδιάδοση με αυξητική μάθηση. Όσον αφορά τους υπόλοιπους αλγορίθμους που συμμετείχαν σε αυτή την έρευνα μπορούν να ειπωθούν τα εξής συμπεράσματα:

- Η οπισθοδιάδοση με ομαδοποιημένη μάθηση έγινε ένας πολύ πιο αποδοτικός αλγόριθμος με μία σημαντικά μικρότερη ποσότητα διαίρεσης των βαρών από τον αριθμό των παραδειγμάτων.
- Η γρήγορη διάδοση έδειξε απογοητευτικά αποτελέσματα ανεξαρτήτως δεδομένων και δομής του δικτύου.
- Η ελαστική διάδοση έδωσε τα δεύτερα καλύτερα αποτελέσματα από τους αλγορίθμους ομαδοποιημένης μάθησης αν και οι παράμετροι που χρησιμοποιήθηκαν για το μέγιστο βάρος ήταν σημαντικά μικρότεροι από αυτούς που προτείνονται στη βιβλιογραφία.

Μία πιο προσεκτική ανάγνωση των αποτελεσμάτων όμως, δείχνει ότι υπάρχουν εναλλακτικές τακτικές, όπως η σταθερή διάδοση. Η σταθερή διάδοση είναι ένας απλός αλγόριθμος με πέντε παραμέτρους, οι τρεις εκ των οποίων είναι ιδιαίτερης σημασίας. Οι παράμετροι αυτοί είναι ο ρυθμός μάθησης και η ορμή που είναι ήσσονος σημασίας, αλλά και οι πιο σημαντικές όπως ο διαιρέτης βαρών, το θετικό – αρνητικό βήμα αλλά και η παράμετρος της συνάρτησης ενεργοποίησης. Σχετικά με το βήμα μπορούμε να πούμε ότι υπάρχουν δύο επιλογές, αυτές του ίσου θετικού και αρνητικού βήματος και αυτή του διπλάσιου θετικού βήματος με προτεινόμενες τιμές τις 1.6 και 0.8. Η σταθερή διάδοση τις περισσότερες φορές έχει συγκρίσιμα αποτελέσματα με τις υπόλοιπες μεθόδους κατηγοριοποίησης, τη στιγμή μάλιστα που τα αποτελέσματα που χρησιμοποιήθηκαν στους συγκριτικούς πίνακες δεν έχουν αυτό το στόχο, μιας και κάτι τέτοιο θα προϋπέθετε μία αποδοτικότερη δομή του δικτύου για κάθε πρόβλημα. Σαν προτάσεις βελτιστοποίησης της οπισθοδιάδοσης προτείνεται ένα σύστημα το οποίο θα είναι ικανό με κάποιο αποδοτικό τρόπο να μειώνει το ρυθμό μάθησης αυξάνοντας το διαιρέτη βαρών. Η ιδιότητα αυτή θα έδινε την ικανότητα στην οπισθοδιάδοση να ψάξει αναλυτικότερα μετά από έναν

αριθμό εποχών εκπαίδευσης.

Αναφορές

[1] Sebastini, F. (2002, March). Machine Learning in Automated Text Categorization. ACM Computing Survets (CSUR), 34(1), 1-47.

[2] Jacobs, R. A. (1988). Increased Rates of Convergence Through Learning Rate Adaption. Neural Networks, 1, 295-307.

[3] Yang, Y., & Liu, X. (n.d.). A re-examination of text categorization methods. 22nd Annual Interational ACM SIGIR conference on Research and Development in Information Retrival, (σσ. 42 - 49).

[4] Krose, B., & Smagt, P. v. (1996). An Introduction to Neural Networks (8th εκδ.). University of Amsterdam.

[5] Bishop, C. M. (1995). Neural Networks for Pattern Recognition. Birmingham , UK: Clarendon Press Oxford.

[6] Rumelhart, D., Hinton, G., & Williams, R. (1985). Learning Internal Representations by Error Propagation. Στο D. Rumelhart, & J. McClelland, Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol 1: Foundations (Τόμ. 1). Cambridge: Bradford Books/MIT Press.

[7] Russell, S. J., & Norvig, P. (1995). Artificial Intelligence A Modern Approach. Prentice Hall.

[8] Fausett, L. (1994). Fundamentals of Neural Networks. Englewood Cliffs, NJ: Prentice Hall.

[9] Hertz, J. A., Krogh, A. S., & Palmer, R. G. (1991). Introduction To The Theory Of Neural Computation. Addison-Wesley Publishing Company.

[10] Fahlman, S. E. (1988). An Empirical Study of Learning Speed in Back-Propagation Networks. Tecnical.

[11] Riedmiller, M., & Braun, H. (1993). A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. IEEE Internation Conference on Neural Networks, (σσ. 586-591). San Francisco, CA

ΔΙΠΛΩΜΑΤΙΚΕΣ ΕΡΓΑΣΙΕΣ

Blerina P. Lika

b.lika@di.uoa.gr

A Novel Approach for alleviating the Cold Start Problem in Recommender Systems

National and Kapodistrian University of Athens, Department of Informatics and Telecommunications
Panepistimioupolis, Ilissia, 15784, Athens, Greece

Abstract

Recommender Systems (RSs) provide personalized suggestions to users for specific items (e.g., books, music) according to their preferences. Popular techniques for building an RS are Content-based (CB) models and Collaborative Filtering (CF). A well-known problem of these systems is the cold start problem that arises when there is lack of information about users or items. In such cases, the RS is not capable of making accurate recommendations. In this thesis, we try to alleviate the discussed problem and propose a hybrid model that incorporates classification methods in a pure CF system by using user demographic data. Through the proposed mechanism, we identify users with similar behavior and predict their ratings for specific items. We evaluate our algorithm and show its performance by executing a large number of experiments with a real dataset (i.e., MovieLens).

Keywords: Recommender systems, cold start problem, Collaborative Filtering, classifiers, semantic similarity

Supervisor

Stathes Hadjiefthymiades, Associate Professor

1. Introduction

In recent years, RSs have become extremely common and are used by several applications and websites such as Amazon, Youtube, IMDB, MovieLens, LinkedIn and Facebook. RSs are information filtering systems that deal with the delivery of information that the user is likely to find interesting or useful. Such systems can make recommendations that are based on models built from item or user characteristics and social environment. RSs encounter a common problem, known as the cold start problem [3]. This problem is detected when the users or items are new to the system and there is no sufficient information (i.e., ratings) in order to provide accurate recommendations. The problem is more intense in the Collaborative Filtering (CF) approaches that are mostly based on user ratings in order to find users similarities and, finally, make recommendations.

Several research efforts tried to propose solutions on the cold start problem. Hybrid approaches that use both CB and CF features are proposed as a solution to the discussed problem [1], [2]. These studies adopt a single probabilistic framework in order to unify CF and CB techniques in sparse data environments. Respectively, in [3] is presented a hybrid approach that is based on the analysis of two probabilistic aspect models that use pure CF and user information. Furthermore, predictive feature-based regression models [4] that leverage all the available information of users and items can tackle the cold start problem. In [5] is proposed a solution adopting the functional matrix factorization (fMF) technique. fMF constructs a decision tree from the initial user interview (each node being an interview question) enabling the RS to query the user adaptively. Hence, the interview phase could be 'alleviated' in order to improve the performance. Finally, the authors in [6] study a novel model for the profiling of the new users and adopt an interview to elicit user opinions on specific items.

In this thesis, we propose a hybrid approach that combines demographic data and CF features in order to alleviate the cold start problem. The proposed algorithm includes three phases: a) user classification, b) user similarity, and, c) rating prediction. The user classification process classifies the new user in a specific group. For this purpose, we adopt widely known classifiers such as C4.5 and Naive Bayes. In the second phase, we utilize an intelligent technique in order to find the neighbors of the new user inside the group. We examine important characteristics of the new user and try to find others inside the group that best matches to her. In the final phase, we predict the ratings of the

new user based on the outputs of previous phases by using a weighted average scheme.

2. Proposed approach

We developed a three-phase algorithm that combines demographic data and user similarity methods on a traditional CF system in order to make accurate prediction on new user ratings. In order to describe in details the proposed approach we define as $U=\{u_1, u_2, ...,u_m\}$ the set of registered users in the RS, $N=\{n_1, n_2, ...,n_n\}$ the set of new users and $I=\{i_1, i_2, ...,i_k\}$ the set of available items. Figure 1 depicts the main components of the proposed approach.

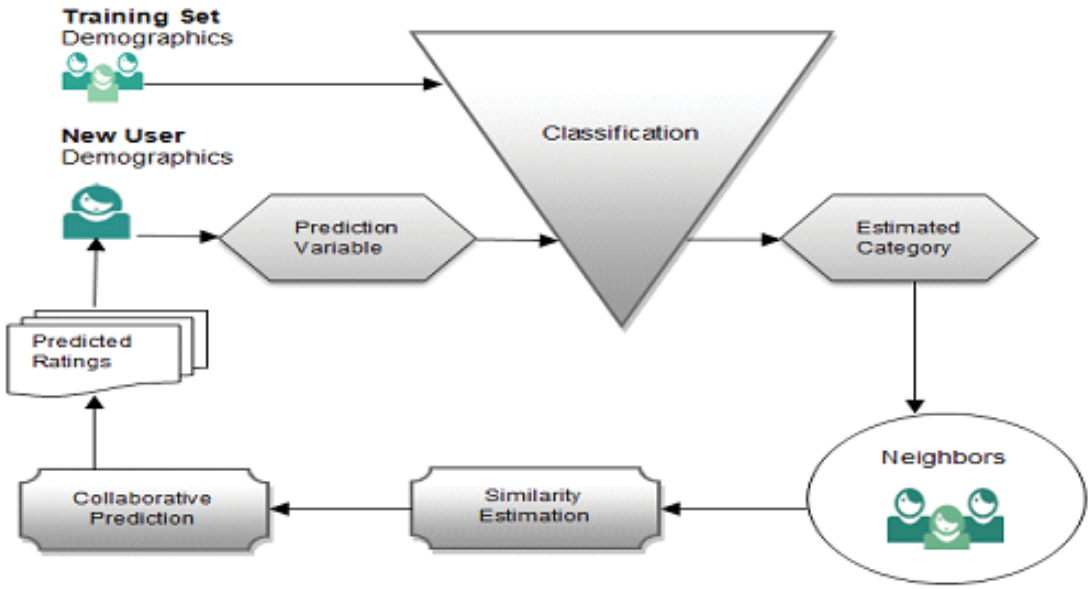


Figure 1: Architectural components of the proposed approach

In the first phase, we use a classifier to build a model based on demographic data (e.g., age, gender, occupation), $D=\{d_1, d_2, ...,d_l\}$ represents the set of each user demographic data. This classification component trains a dataset that contains instances (observations) with D attributes. The training process generates a model that maps a new user in the appropriate category according to her demographic data. Let $C=\{c_1, c_2, ...,c_b\}$ be the set of possible categories that the user may belong. Figure 1 shows two key factors: a) the prediction variable V, and, b) the estimated category $\tilde{C}$. For each

new user we set a class attribute that represents V. The values of this class are the possible categories $\bigcup c_j \subseteq C$ in which we will classify the new user. One of these categories c_j is the output of the model and the corresponding category $\tilde{C}$ for every $n \in N$. The goal of this process is to find a neighborhood $NG = \bigcup u_j (NG \subseteq U)$ for every $n \in N$. The neighbors in NG are users that belong to the same category predicted by the model.

After the definition of NG, we calculate the similarity between $n \in N$ and each of the neighbors $u_j \in NG$, $j=1,2,...,|NG|$ through a weighted average of their demographic data. More specifically, we obtain the similarity for the age, gender and occupation and we combine these three metrics in a weighted scheme producing an overall weight for the similarity between the new user and her neighbor. This way, we are able to predict ratings of the new user based on top-k nearest neighbors. In case of numeric data (e.g., age), we adopt an exponential function that is described in Section 2.2. For nominal values (e.g., occupation), we adopt the Wu-Palmer semantic similarity metric [8].

Finally, predictions are produced based on a weighted sum of NG ratings. We combine similarities retrieved by the previous steps with the ratings of NG. In this step we implement a function that makes a prediction for an item $i \in I$.

2.1. User classification

Through the adoption of classification algorithms, we produce the estimated category $\tilde{C}$ based on the data related to U. We apply a multiclass classifier in order to generate multiple categories. More specifically, we turn a binary classifier into a multinomial classifier using the one against all (one-vs-all) strategy [7]. For each class, we train a single classifier in order to distinguish that class from all other classes. The final predicted class is that with the highest confidence score that satisfies Equation (1).

$$\hat{y} = argmax_{(1 \le k \le K)}^{f(x)} \qquad (1)$$

Figure 2 depicts the OvA algorithm that we use in order to implement a multiclass classifier.

Algorithm 1. OvA

Input: L, Instances X, Labels Y! L: Training Algorithm
Output: Classifiers f_k, $k = 1, 2, \ldots, K$
Begin
  forall $k \in \{1, 2, \ldots, K\}$ **do**
    if $y_i = k$ **then**
      $y'_i = 1$
    else
      $y'_i = 0$
    end if
    $f_k = L(X, y')$
  end for
End

Figure 2: Multiclass classification (OvA)

2.2. User similarity

After retrieving $\tilde{C}$, we group users according to their categories. Our aim is to find the neighborhood for each new user $n_j \in N$. Figure 3 shows the algorithm that calculates user neighborhood. It matches C_{n_j} against C_{u_j}, where C_{n_j} is the category of the new user $n_j \in N$ and C_{u_j} is the category of the user $u_j \in N$. The result is the set of neighbors NG. Figure 3 presents the algorithm that calculates NG group.

Algorithm 2. Neighborhood Calculation

Input: $\mathcal{U}, \mathcal{N}$
Output: $\mathcal{NG}$ for each new user
Begin
  Define options: OvA
  Set Class Index
  Build C4.5 Tree
  Build the MultiClass Classifier
  for all $n_j \in \mathcal{N}$ **do**
    $\mathcal{NG} = null$
    Find $\tilde{C}_{n_j}$
    for all $u_j \in \mathcal{U}$ **do**
      Find $\tilde{C}_{u_j}$
      if $\tilde{C}_{n_j} = \tilde{C}_{u_j}$ **then**
        $\mathcal{NG}$.add(u_j)
      end if
    end for
  end for
End

Figure 3: Algorithm for new user neighborhood calculation

In addition, we calculate the weight of the similarity on the demographic data between the new user and NG through Equation (2).

$$sim(n,u) = \sum_{j=1}^{l} SF_J \cdot w_j \qquad (2)$$

In Equation (2), SF_j is the similarity value of the j^{th} attribute (e.g., similarity of age) and w_j is the corresponding weight. In this way, we can pay more attention on specific demographic data. For example, let us consider $D=\{d_1=age,\ d_2=occupation,\ d_3=gender\}$ so that $l=3$. We can pay more attention on age, if we define $w_1=0.5,\ w_2=0.25,\ w_3=0.25$. In order to calculate the similarity value SF_j for each attribute d_j, we define a similarity function $SF(at_1,at_2) \in [0,1]$. The terms at_1 and at_2 are the attribute values that will be compared for a pair of users. We consider two attribute categories: (a) numeric, (b) nominal. For numerical values we use an exponential function $SF : R^+ XR^+ \rightarrow [0,1]$ defined in Equation (3):

$$SF = \begin{cases} (1 - \frac{|Diff|}{Diff_{\max}})^{\omega} & if\ |Diff| \le Diff_{\max} \\ 0 & if\ |Diff| > Diff_{\max} \end{cases} \quad (3)$$

More specifically, $Diff$ is the age difference between two users and $Diff_{max}$ is a maximum difference (defined by developers). The ω parameter is a policy factor. For nominal values we adopt the Wu and Palmer semantic similarity metric. Wu-Palmer metric adopts the known Least Common Subsummer (LCS). This technique results the common node of the values to be compared according to the Wordnet taxonomy[1]. Finally, in case of binary nominal attributes (i.e., gender) or binary numerical attribute values, we consider boolean similarity values (true or false). Hence, $SF(at_1, at_2)=1$ *when* $at_1=at_2$ *and* $SF(at_1, at_2)=0$ *when* $at_1 \neq at_2$.

2.3. Rating Prediction

The final phase is the prediction of ratings for specific items. For each new user $n_j \in N$, the model aims to provide predicted ratings for every item $i_b \in I$. Every predicted rating $R_{n_j,i_b} \in R^+$ is calculated by the weighted sum of NG ratings for specific items $i_b \in I$ (Equation (4)):

$$R_{n_j,i_b} = \frac{\sum_{u \in NG} sim(n_j,u) \cdot r_{u,i_b}}{\sum_{u \in NG} sim(n_j,u)} \quad (4)$$

In Equation (4), r_{u,i_b} is the rating of the user u (neighbor of n_j) for the item i_b. Based on this approach, we aim to enhance ratings that are made by users having large similarity with the new user.

1. http://wordnet.princeton.edu/

3. Experimental Evaluation

We evaluate the performance and prediction accuracy of the proposed approach based on the widely known metrics Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

$$MAE = \frac{1}{K}\sum_{u,i}\left|p_{u,i} - r_{u,i}\right| \quad (5)$$

$$RMSE = \sqrt{\frac{1}{K}\sum_{u,i}\left(p_{u,i} - r_{u,i}\right)^2} \quad (6)$$

We run several experiments using the MovieLens[2] dataset containing one million ratings for 4000 movies provided by 6000 users. From the set of users, we choose a number as the registered users in the system and the rest are considered as new users. We start from 100 registered users and, in different experimental scenarios we increase the number till 5000 users. Through this approach, we try to find out how the system behaves for different registered users number. Ratings are between 1 (minimum value) and 5 (maximum value). All ratings are considered to be integer values. We adapt the MovieLens dataset at the proposed model by adding in the dataset a category that the user may belong according to her movie preferences. Therefore, we consider four attributes $C=\{c_1, c_2, c_3\}=\{fun, intellectual, adventurous, romantic\}$. Hence, a user is defined by id, her demographic data $D=\{d_1, d_2, d_3\}=\{age, occupation, gender\}$ and the respective category $\tilde{C}$. Moreover, in the classification step we use two binary classifiers: a) C4.5 and b) Naïve Bayes in order to build a multiclass classifier. We compare our results of the proposed algorithm when adopting these classifiers with those when users are randomly classified in $\tilde{C}$. For this purpose, we replace the classifiers in the first step of the proposed approach with a Random Classification Algorithm (RCA). For the C4.5 approach, we examine a scenario where only two classes are used for the classification of each user (C^2 4.5) and a scenario where multiple classes are considered in the classification process (C^M 4.5). Table 1 depicts an overview of our experimental

2. http://grouplens.org/datasets/movielens/

parameters.

Parameter	Values
Algorithm	$C^2$4.5, C^M4.5, NB, RCA
w_j	$w_j \in [0,1], \sum_{j=1}^{3} w_j = 1$
ω	0.8

Table 1: Experimental parameters

We examine different scenarios defined by the weights w_j values for each d_j. Table 2 depicts four scenarios. Every combination focuses on a specific combination of weights for demographic data. For instance, in Scenario 1, the algorithm pays more attention on the age in order to make the necessary recommendations. Scenario 4 is more "fair" as all the demographic data are equally considered for recommendations.

Scenarios	Weights
Scenario 1	$w_1 = 0.6, w_2 = 0.3, w_3 = 0.1$
Scenario 2	$w_1 = 0.3, w_2 = 0.6, w_3 = 0.1$
Scenario 3	$w_1 = 0.3, w_2 = 0.1, w_3 = 0.6$
Scenario 4	$w_1 = 0.33, w_2 = 0.34, w_3 = 0.33$

Table 2: Experimental scenarios

Figure 4 shows results of MAE and RMSE when we consider the Scenario 1. For both metrics, C^2 *4.5* algorithm has the best performance. As $|U|$ (number of users) grows, MAE and RMSE are reduced. For $|U| = 900$, we take MAE approximately equal to 0.8 and RMSE approximately equal to 1.0. As $|U|$ increases, the system has more data to achieve good performance in the classification process as well in matching user demographic information. Hence, the prediction error becomes smaller compared to the scenarios with few users. The respective results are shown in Figure 5. In this case, the best performance is achieved by the C^M *4.5* algorithm accompanied by the NB. The minimum MAE value is equal

to 0.736 achieved by $C^M\ 4.5$ when $|U| = 5000$. As natural, the RCA algorithm performs worse than the rest. We obtain similar performance for the rest of the examined scenarios.

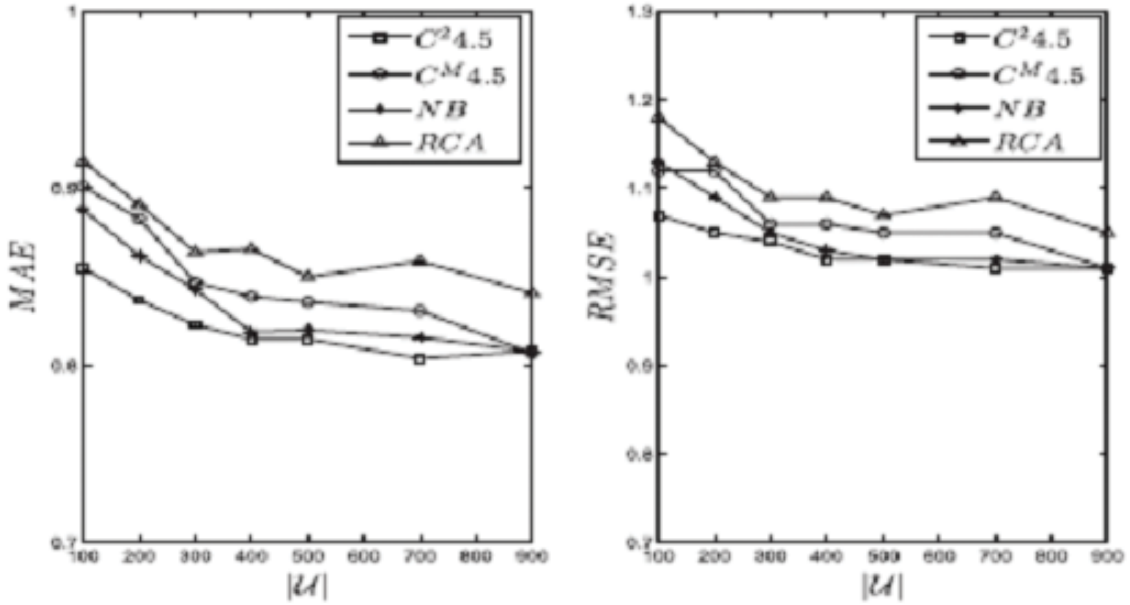


Figure 4: Results for Scenario 1

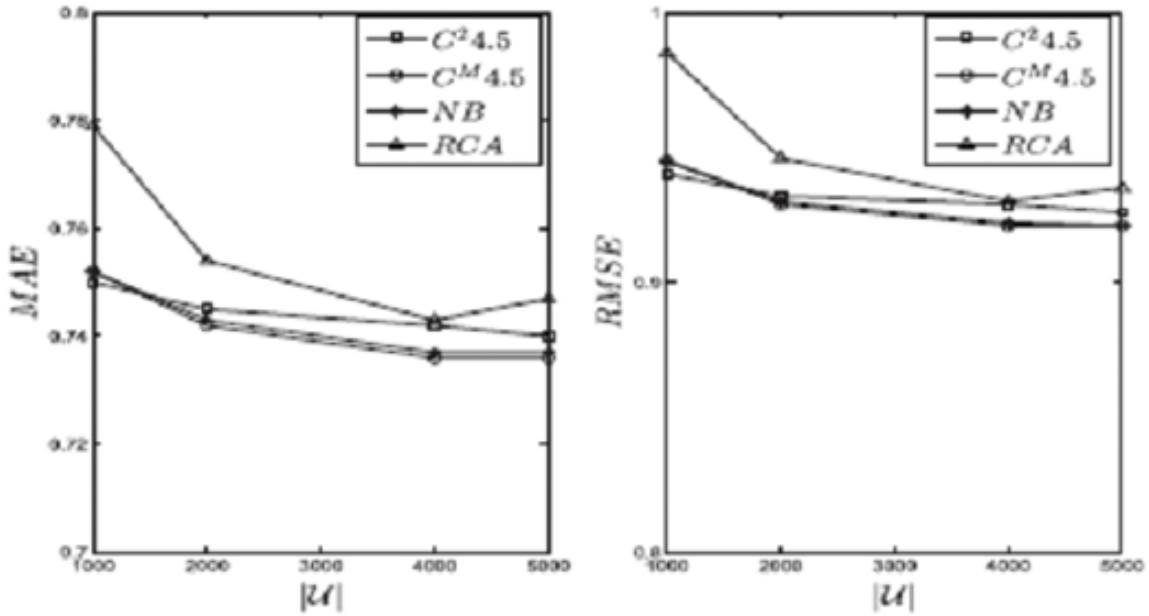


Figure 5: Results for Scenario 2 - large number of users

In order to compare the performance of the proposed approach using a small dataset (e.g., $|U| \in \{100, 200, 400, 500\}|$) as well using a large dataset (e.g., $|U| \in \{1000, 2000, 4000, 5000\}|$) we define the D metric (Equation (7)).

$$D = \frac{D_{Base} - D_{Target}}{D_{Base}} \% \qquad (7)$$

More specifically, D_{Base} stands for $Base \in \{100, 200, 400, 500\}$ and D_{Target} for $Target = 10 \cdot Base$. We calculate the D_{MAE} and D_{RMSE} for both MAE and RMSE metrics. Figure 6 shows the results for D_{MAE}. We see that all the algorithms are affected by the increase in $|U|$. The C^2 4.5 algorithm is less affected compared to the rest. The difference in the performance becomes smaller as $|U|$ increases. However, the difference remains close to 10% as Base = 500. Concerning the RMSE metric, we see that the C^M 4.5 algorithm is heavily affected by the increase in $|U|$ as in the MAE case. Smaller $|U|$ leads to greater MAE and RMSE results. This is because the system does not have enough information about users in order to get better predictions.

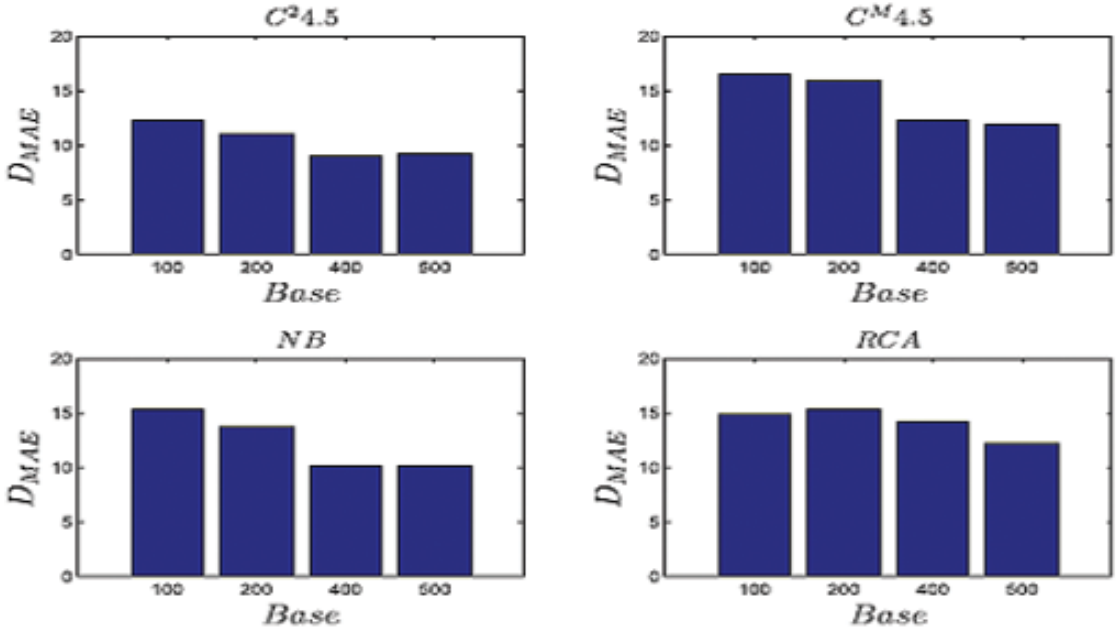


Figure 6: D_{MAE} comparison

4. Conclusions

The proposed approach addresses the new user cold start problem for RSs applying pure collaboration filtering techniques. We adopt a three-phase algorithm in order to make predictions about new user's rating for an item. We combine classifiers and similarity techniques for defining users with similar characteristics. The idea is that people with a common background and similar features might have similar preferences for items. The new user is classified in a group according to her demographic data. Users of this group are her neighbors whose ratings are combined in a weighted scheme in order to calculate the predicted rating of the new user. Experimental results indicate that the proposed approach can efficiently provide accurate recommendations in the absence of any information about users. As a result, the proposed method performs better in scenarios where decision trees are used in the classification phase as well for a large number of registered users.

References

[1] A. I. Schein, A. Popescul, L. H Ungar, and D. M. Pennock, "Methods and metrics for cold- start recommendations", In Proceedings of the 25th international ACM SIGIR, 2002.

[2] Popescul, L. H. Ungar, D. M. Pennock, and S. Lawrence, "Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments", In Proceedings of the 17th conference on uncertainty in artificial intelligence, 2001.

[3] X. N. Lam, T. Vu, T. D. Le, and A. D. Duong, "Addressing cold-start problem in recommendation systems", In Proceedings of the second international conference on ubiquitous information management and, communication, 2008.

[4] S. T. Park, and W. Chu, "Pairwise preference regression for cold-start recommendation", In Proceedings of the third ACM conference on recommender systems, 2009.

[5] K. Zhou, S. H. Yang, and H. Zha, "Functional matrix factorizations for cold-start recommendation", In Proceedings of the 34th international ACM SIGIR, 2011.

[6] N.Golbandi, Y.Koren, and R. Lempel, “Adaptive bootstrapping of recommender systems using decision trees”, In Proceedings of the 4th international conference on web search and web data mining, 2011.

[7] J. Milgram, M. Cheriet, and M. Sabourin, “One against one or one against all: Which one is better for handwriting recognition with SVMs?”, In 10th International workshop on frontiers in handwriting recognition, 2006

[8] Z. Wu and M. St. Palmer, “Verb Semantics and LexicalSelection”, Association for Computational Linguistics (ACL), 1994, pp. 133-138

Extended version of this thesis is published in Expert Systems with Applications, Elsevier 41(4): 2065–2073 (2014), with the title “Facing the Cold Start Problem in Recommender Systems”

Georgios P. Nomikos

gnomikos@di.uoa.gr

Point centrality indices and ISP network vulnerability

National and Kapodistrian University of Athens, Department of Informatics and Telecommunications
Panepistimioupolis, Ilissia, 15784, Athens, Greece

Abstract

The position of the nodes within a network topology largely determines the level of their involvement in various networking functions. Yet numerous node centrality indices, proposed to quantify how central individual nodes are in this respect, yield very different views of their relative significance. Our first contribution is then an exhaustive survey and categorization of centrality indices along several attributes including the type of information (local vs. global) and processing complexity required for their computation. We next study the seven most popular of those indices in the context of Internet vulnerability to address issues that remain underexplored in literature so far. First, we carry out a correlation study to assess the consistency of the node rankings those indices generate over ISP router-level topologies. For each pair of indices, we compute the full ranking correlation, which is the standard choice in literature, and the percentage overlap between the k top nodes. Then, we let these rankings guide the removal of highly central nodes and assess the impact on both the connectivity properties and traffic-carrying capacity of the network. Our results confirm that the top-k overlap predicts the comparative impact of indices on the network vulnerability better than the full-ranking correlation. Importantly, the locally computed degree centrality index approximates closely the global indices with the most dramatic impact on the traffic-carrying capacity; whereas, its approximative power in terms of connectivity is more topology dependent.

Supervisors

Ioannis Stavrakakis, Professor | Merkourios Karaliopoulos, Senior Researcher | Panagiotis Pantazopoulos, PhD candidate

1. Introduction

Social Network Analysis (SNA) provides a highly interdisciplinary theoretical framework for processing social information and analyzing social structures [1]. It draws heavily on graph models mapping individual actors within the social network to the graph vertices and their relationships to the graph (weighted) edges. It then leverages graph-theoretic concepts, metrics and results to answer questions about the relative importance of actors and the way information flows across it. Centrality is one such concept/metric. To the best of our knowledge, it dates back to the work of Bavelas [2], who first gave a formal definition of node centrality in connected graphs as the sum of its geodesics (shortest-paths) to all other nodes. New indices were proposed and existing ones were adapted to apply to a broader range of scenarios [4].

Motivation and objectives: Our main objective in this study is to quantify how much information is embedded in centrality indices about the relative importance of Internet nodes for different network operations. Given that all centrality formulations proposed in literature are heuristic, the questions that naturally arise are how do these formulations compare in their assessments/ predictions about the nodes' relative importance and which one(s) may be the "right one(s)" to consider as reference for more reliable predictions of network vulnerability. The thesis seeks to systematically address these questions by undertaking a three-step study with various instances of methodological innovation. The first step, which herein is briefly presented, involves a thorough survey and novel classification of the variety of centrality indices proposed in literature over the last sixty years. This classification is then used to select the seven most popular and representative indices for carrying out the two experimental steps of the study. Hence, as a second step, we derive the node rankings these indices induce over more than 40 router-level

snapshots of network topologies and study their correlation. The correlation strength is assessed by the mainstream rank/linear correlation coefficients, but also less widespread measures such as the percentage overlap in the lists of the k most central nodes. Finally, we compare the seven indices with respect to their capacity to reveal the network vulnerability to node removals. Hence, we let the indices dictate the most central nodes to-be-removed and assess how the network connectivity properties but also its traffic-carrying capacity are affected.

2. A Novel Classification of Centrality Indices

In this Section, we briefly present the way we have characterized and classified the rich variety of centrality indices that we have run across in our study of the highly interdisciplinary 60-year-old literature [4]. At a first-level the reviewed indices are divided into node (point) centrality and graph centrality indices. The former are addressed by the vast majority of the literature and concern individual nodes; whereas the latter are derived for whole graphs as functions of the individual node centrality indices. Then, node centrality indices are further characterized using three fundamental attributes, briefly discussed next:

Centrality context: topological vs. flow-aware. The vast majority of centrality indices takes only the network topology into account. They reflect either the distance of a node from all other network nodes or the extent to which a node lies on paths connecting other network nodes [3]. Topological centrality indices also include the so-called spectral indices, which depend on the eigenstructure of a matrix (e.g., adjacency or Laplacian) related to the network in question. The second set groups indices, which attempt to factor the (predicted) network traffic in the centrality computation [4].

Underlying graph types. Most of the indices are defined over connected, undirected, binary, static graphs. Efforts to relax in turn each one of these four graph attributes have resulted in a plethora of indices that can cope with disconnected, directed, weighted and dynamic types of graphs [4].

Computational Aspects. Centrality indices can be separated into local and global ones, depending on the extent of topological information that is required to compute them. To limit the scope of centrality computations one may use the sociological notion of the ego-network [5] or control the length k of the considered paths [6].

Selecting centrality indices for experimentation. Out of the numerous indices reviewed in [4], we select the seven most popular ones that appear repeatedly in the literature and, at the same time, capture a wide range of different notions of centrality. Those indices are the Degree (DC), Betweenness (BC), Closeness (CC), Eigenvector (EC), Harmonic Centrality (HC), Pagerank (PG, with d=0.85 as typically used in literature) and Eccentricity (ECC) [3][4].

3. Correlation Study Of Centrality Indices

In almost all instances where centrality indices inform network protocols, what matters is the ranking of nodes induced by those indices rather than their absolute values. These rankings are subsequently used in the decisions made by the respective protocols. The question that plausibly arises in every case is how similar are the rankings generated by each centrality index. In this section, we carry out a thorough correlation study of these rankings, computed over a broad set of ISP router-level topologies. First, we calculate for each topology and node in it the seven centrality indices (Section II), thus generating seven different node rankings per topology. Then, we compute pairwise correlation measures over these rankings. We consider two different measures, one accounting for the full node rankings and the other only for the most highly-ranked nodes.

3.1. Index correlation measures and router-level topologies

Index correlation measures. The first correlation measure is the nonparametric Spearman's rank-correlation coefficient, ρ_V, and is computed over the full node rankings. For a given network topology node set V, it is:

$$\rho_V(C_1, C_2) = 1 - \frac{6 \sum_{u \in V} (r_{C_1}(u) - r_{C_2}(u))^2}{|V|(|V|^2 - 1)}$$

where r_{C1} (u) and r_{C2} (u) are the ranks of node u in line with centrality indices C_1 and C_2, respectively. It lies in [−1, 1], with high positive (negative) values denoting strong positive (negative) correlation. The second correlation measure is the percentage overlap between the sets of the k most highly ranked (top-k) nodes that are generated by two indices.

$$ov_V(C_1, C_2; k) = \frac{|\{v \in V : r_{C_1}(v) \leq k\} \bigcap \{v \in V : r_{C_2}(v) \leq k\}|}{k} \cdot 100\%$$

The relevance of the two measures depends on the usage context of centrality-based ranks.

Router-level ISP topologies. All our experiments are carried out over four datasets, Rocketfuel [7], CAIDA [8], mrinfo (Tier-1 and Transit) [9] (binary router-level graphs) and Topology Zoo (capacitated topologies) [10].

3.2. Results

Full-ranking correlation over binary graphs. Based on our results, the first remark is that not a single centrality pair is negatively correlated over any of the studied topologies. We empirically characterize the pairwise index correlation as high and low when the corresponding ρ_V values lie in the intervals [0.7,1] and [0.3,0.7), respectively. On the other hand, two indices are considered non-correlated when their ρ_V lies in [0-0.3). The second point is that the indices' correlation values follow similar trends across all datasets so that they can summarized graphically in a graph like the one of Fig. 1.a. No edges are added for non-correlated index pairs.

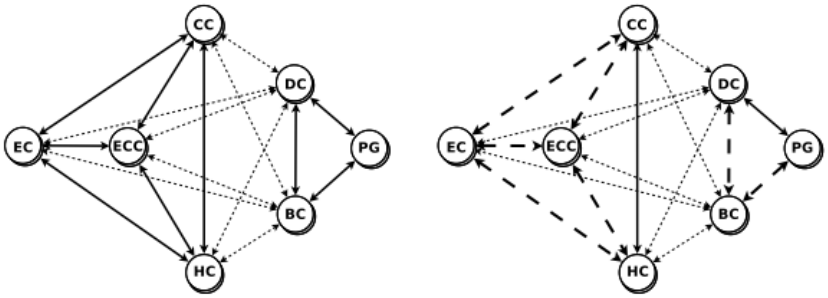


Fig. 1. Graph-based illustration for the average values of the Spearman coefficients (a) and top-5% overlap (b) among centrality indices. In (a) solid bold and dashed plain lines denote coefficients in the intervals [0.7-1],[0.3-0.7), respectively. In (b), solid bold, dashed bold and dashed plain lines denote overlap higher than 70%, between 40-70%, and lower than 40%, respectively.

Betweenness vs. Degree centrality: Degree centrality (DC) captures, at least phenomenally, a completely different notion of centrality than Betweenness (BC). DC takes into account only the node's local neighbors, whereas BC considers the position of the node within the whole network. Therefore, in some cases DC can evaluate nodes' position very differently than BC; In our datasets, the two indices are found consistently highly correlated, in agreement with earlier studies [5], [11], [12] that report positive Pearson correlation between DC and BC over a wide range of networks such as random graphs and real-world complex networks.

Pagerank vs. Degree centrality: Another persistent result, immediately apparent from Fig. 1.a, is the strong correlation between Pagerank (PG) and DC. PG is principally defined for digraphs discriminating between incoming and outgoing connections at each node. Taking into account the aforementioned strong BC-DC correlation, a triangle-like schema emerges and may be of practical importance as it relates DC, the only local, index with two globally-determined ones. Grolmusz shows in [13] for undirected general graphs that Pagerank is statistically close but not identical to the degree distribution. Positive correlation between the three indices (PG-DC-BC), with ρ values in [0.66, 0.95] for all three-index pairs, is also reported in [14] over co-authorship real-world data (directed graphs).

Pagerank vs. Eigenvector centrality: PG, and EC centrality are the two spectral indices we experiment with. Both express the stationary probability of

a random surfer to reside on some page while moving on the Web graph. Hence, one would expect some positive correlation between these indices. However, our results indicate the absence of such a relationship. A possible cause is that their actual interpretation differs as, contrary to EC, the PG centrality utilizes the damping factor d to determine the "jump" probability.

Eccentricity vs. Closeness centrality: Another strong correlation is between the ECC and CC centrality indices. Recalling the definitions of the two indices, there is absolute positive ECC-CC correlation if it holds that $ECC(n_1)>ECC(n_2)$ whenever $CC(n_1)>CC(n_2)$, for all $n_1, n_2 \in V$. We can rewrite the former equation as $max_{j\in V}d_{n2,j}>max_{j\in V}d_{n1,j}$ and the latter as $\Sigma_{j\in V}d_{n2,j}>\Sigma_{j\in V}d_{n1,j}$. Hence, looking at the last two inequalities, the question becomes when the order in maximum index values is also preserved for their averages over the studied graph. This holds in several trivial graphs (e.g., line graph, rectangular grid) but not in all graphs.

Top-k percentage overlap over binary graphs. So far, our correlation analysis has taken into account the full rankings produced by the seven centrality indices. We now focus our attention on the top-5% most central nodes identified by each index and investigate how large are the overlaps between different rankings. The motivation for this set of experiments is the existence of network protocols that seek to exploit a small set of the most central nodes [4]. In Fig. 1.b we show a summarizing graph-based illustration of the overlap scores among the seven centrality indices. Figure 4 presents the average overlap of nodes over all ASes of each dataset for the most significant centrality pairs. On the one hand the overlap of some indices (e.g., BC-CC or HC-BC) appear to be highly sensitive to the considered topology, with differences that reach 40% across different datasets. On the other, all pairs found earlier to be strongly correlated in terms of full rankings, appear to be more weakly associated in terms of overlap values. This result should come as no surprise since rank correlation is determined over all network nodes rather than a subset of cardinality k. Let us look closer into the BC-DC pair. Fig. 5 illustrates how the number of nodes with DC=1 affects the rank correlation coefficient. It seems that the Spearman values between the two indices increase with the number of DC=1 nodes. These nodes are expected to positively contribute to the DC-BC correlation as they also exhibit the lowest-ranked betweenness value (i.e., BC=0). At the same time, the ones

with the top BC and DC values may not necessarily coincide. The above results suggest that the high DC-BC correlation is mainly due to nodes of lowest ranks. This observation warns against the actual value of high Spearman rank correlation coefficients between two indices. On the other hand, the overlap measure does not suffer from similar biases.

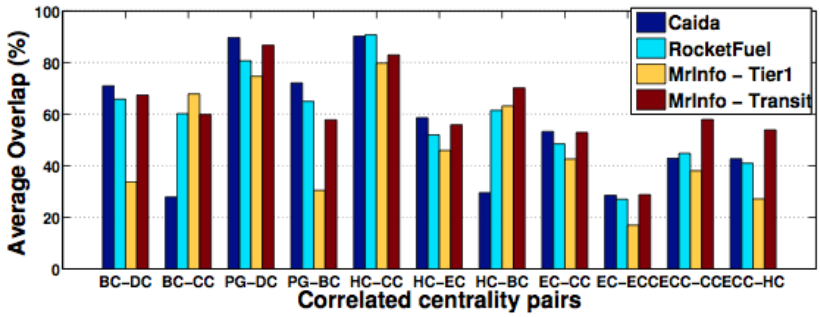


Fig. 4. Mean overlap (%) between the top-5% nodes of centrality rankings.

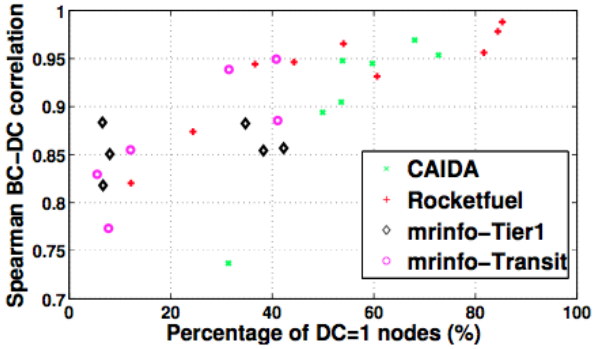


Fig. 5. The relation between the BC-DC rank correlation and the percentage of nodes with degree equal to one.

4. Centrality And Network Vulnerability

The network vulnerability analysis is of interest to various parties. A potential attacker would like to know which index results in node removals with the most significant impact on the network performance, so as to orchestrate the most effective attack. From the network operator's side, the dual aim is to identify and better protect those critical nodes, whose failure would result in

maximum network performance degradation. In this thesis, we relate the term "performance" to fundamental connectivity and traffic capacity properties of the network rather than the scores achieved by specific protocols/applications. This way we get away with their engineering details that shape the end impact and place the emphasis on the network topologies per se.

4.1. Centrality-driven node removals and connectivity

The experiments of this section explore how the size of the giant connected component and the total number of connected components in each topology are affected when up to 5% of the network nodes are removed. The experiments are carried out over the binary datasets described in the subsection III-A.

Size of giant connected component (GCC): The GCC size reflects the number of nodes that can communicate with each other. Figs. 6.a,d suggest that removing those vertices that the ECC index identifies as most central has the minimum impact on GCC. All other indices expose more quickly the vulnerability of the network but we cannot identify any dominance relationship among them that persists over all datasets. However, a closer look reveals that it is the top-k overlap between two indices, rather than their rank-correlation that essentially determines how similar is the impact of the corresponding removals. Overall, a concluding note would be that any two indices measured with high top-k overlap values are expected to give rise to similar GCC sizes, and vice versa. The full-rank correlation values are not always in line with the experienced impact due to the biases discussed in Section III.

Number of connected components: Again, the ECC index yields node removals that result in minimum network fragmentation (Figs. 6.b,e). Interestingly, DC, a purely local index succeeds in removing nodes that play critical role in connectivity as opposed to the other global and more complex ones (except PG). On the other hand, BC and DC which were also found strongly rank correlated yet of weaker top-k overlap, have different impact on the connected components. Removing nodes according to DC, the number of components increases constantly compared to the impact of BC. This implies that the network connectivity mainly relies on strategic hub-nodes rather than bridging nodes that are typically of high BC.

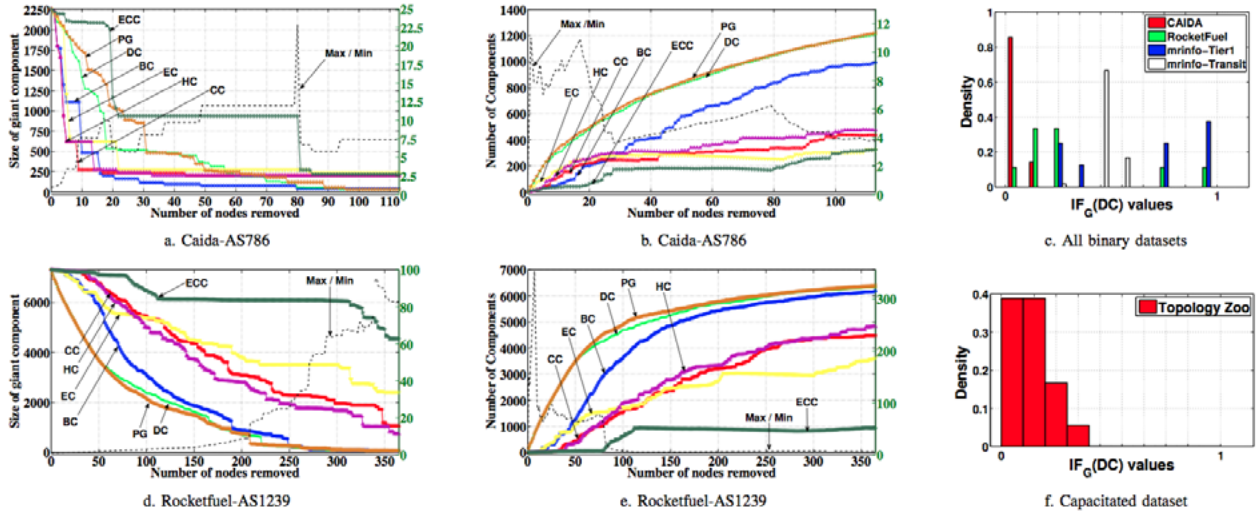


Fig. 6. a,b,d,e) Effect of node removals on the size of the giant-connected component (a,d) and the number of components (b,e) for two indicative ASes. c,f) Empirical probability mass function of the $IF_G(DC)$ measured *w.r.t* the size of the giant component (c) and the max flow accommodated by Topology Zoo (f).

Local vs. global centrality indices: Figures 6.a,b,d,e clearly show that the removal of the most central nodes affects differently the connectivity measures depending on which centrality index is used to determine them. For each number k of removed nodes, one can identify best- and worst-case values, $m_{bc}(k)$ and $m_{wc}(k)$ respectively, for the two performance metrics. These values may be obtained by different centrality indices as the considered metric m changes. Essentially, we seek to quantify how close to the best-/worst-case is the impact of removals when directed by the single locally computable centrality index. To this end, for each centrality index c, topology G, number of removed nodes k and performance metric $m(k;\ c)$ we define a normalized distance measure, hereafter called impact factor $IFG(k;\ c)$ as:

$$IF_G(k;c) = \frac{|m(k;c) - m_{wc}(k)|}{|m_{bc}(k) - m_{wc}(k)|}$$

Note that depending on the metric, the worst-case value may coincide with the minimum or maximum value the metric gets over all indices. It is then straightforward to derive a topology average measure of the impact factor as:

$$IF_G(c) = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \frac{|m(k;c) - m_{wc}(k)|}{|m_{bc}(k) - m_{wc}(k)|}$$

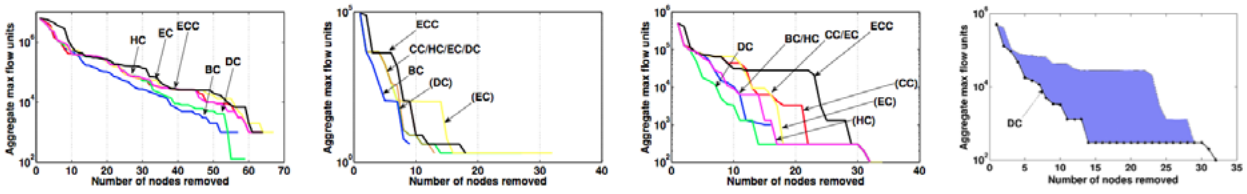
where K is the set of k values considered in the evaluation. Clearly, both $IF_G(k;\ c),\ k \in K$ and $IF_G(c)$ take values in [0, 1]. We are particularly interested in $IF_G(DC)$ and Fig. 6.c plots the empirical probability mass function of the $IF_G(DC)$ values over all topologies of a given dataset, when the metric m is the size of the GCC. Despite its local nature, DC-driven node removals in most cases affect significantly the GCC size. To which extent this impact approximates the worst-case over all indices depends on the underlying topology. Over the CAIDA networks DC closely approximates the low end of the envelope (an envelope plot like the one in Fig. 8.d. indicates the best- and worst- value of a certain connectivity metric for all centrality-driven node removals). The approximation is looser over Rocketfuel, whereas in the mrinfo (Tier-1) and (Transit) networks, considerable mass is accumulated at medium and high IFG (DC) values, respectively.

4.2. Centrality-driven node removals and traffic capacity

We now turn our attention to a much less investigated topic, the comparative impact of centrality-driven node removals on the network traffic serving capacity. Such a task is not straightforward. One approach would be to consider a given traffic matrix utilizing the solution of an instance of the multicommodity flow (MCF) problem. However the MCF problem is an NP-complete problem, with the computational complexity raising fast with the number of commodities. To overcome those limitations, we have taken a simpler approach and estimate the traffic serving capacity of the network as the sum of maximum flows over all network node pairs [4]. Namely, we iterate over all node pairs and for each pair we solve an instance of the maximum flow problem, i.e., compute the maximum traffic load that can be served by the network when only the particular pair transfers traffic across the network. Clearly, this sum is a (very) loose upper bound of the traffic load that can simultaneously be served by the network. However, it provides a traffic load neutral measure of what can the network carry and how is this affected when a variable number of nodes is removed.

Experimentation methodology and results. Our experimental study is carried out over the Zoo Internet topologies with capacitated (weighted) links. For determining the node rakings we had to carry out the centrality indices computations over weighted graphs. This was mainly a question of computing

shortest paths over weighted graphs. Regarding the spectral indices, in the topology Zoo experiments we only employ the EC index that lends to a straight forward extension over the weighted graphs. We then removed nodes in decreasing order of centrality and measured the aggregate maximum flow over all node pairs. The computed aggregate maximum flow over an indicative set of networks is plotted in Figs. 8.a-c. We have obtained similar results for the rest of Zoo datasets (totally 18 snapshots). The rate of aggregate max flow reduction with the fraction of removed nodes varies wildly. This results in high best-to worst-case flow values and wide envelopes, as shown in Fig. 8.d. Highly correlated index pairs, especially those with high top-k percentage overlaps, affect the accommodated flow in similar ways (i.e., intersection of corresponding curves). On a positive note, when node removals are driven by the DC index, the resulting aggregate maximum flow in most cases of Fig. 8.a-c is very close to the worst achieved over all indices. This is more clearly shown in the empirical probability mass function of the IFG (DC) measure in Fig. 6.f, whose mass is highly concentrated at (very) low values close to zero. On the contrary, the considered networks exhibit their highest resilience against the ECC-driven node removals. This behavior can be explained along the same arguments employed earlier, when discussing how node removals affect the connected components. Having a single node i.e., the furthest one, determine ECC may result in some of the most central nodes not being included in the top positions of the ECC ranking.



a. Uninett I_mean b. Carnet c. Bren d. Bren

Fig. 8. a,b,c) Impact of node removals on the maximum flow that Zoo topologies accommodate. When curves coincide, a single identifier points to multiple indices; when they become separated, each one is pointed with its index in parenthesis. d) Envelope plot of the DC-based node removal effects on the max flow.

5. Related Work

Regarding survey studies, Freeman [3] back in 1979 reviewed several centrality indices and much later, Borgatti [16] introduced a typology of the different types of network flows. A graph theoretic review in [6] classifies centrality

measures according to their computational requirements. With respect to centrality correlation studies, we are aware of two works [11][12] that compute linear correlation between the DC and BC indices. Neither of them assesses how the network is affected when different indices are used to direct node removals. Work along this thread typically addresses synthetic graphs and the removals' impact is measured through purely topological measures. Hence, in [16] the scale-free topologies are found vulnerable to the removal of high-degree nodes and in [15] removals of high-DC and -BC nodes in an AS-level topology are found equally harmful in terms of the inverse geodesic length and the number of connected components.

6. Conclusions

We have undertaken a systematic approach to study the relevance of node centrality indices to the Internet vulnerability. Departing from an exhaustive survey and a novel classification scheme of numerous centrality indices, we have carried out a thorough correlation study of the node rankings. Then, we have experimentally assessed the impact of node removals determined by those rankings. Contrary to previous works that consider only network connectivity issues we have extended the vulnerability context to the network traffic-serving capacity. Our main results follow:

- Certain index pairs (such as DC-BC, DC-PG) were consistently found to be high (rank-) correlated across all datasets. Yet a significant part of the high full rank correlation is due to the nodes that are ranked last.
- Node removals based on initial centrality rankings showed that index pairs may exhibit dissimilar impact on the connectivity despite their high (rank-) correlation.
- ECC is consistently the index with the least impact. On the other hand, local-only information (DC) will be practically used to approximate the index with the worst impact. In terms of connectivity, such an approximation depends on the underlying network. In terms of the network traffic capacity, the approximation is highly effective implying that the complexity of global indices can then be escaped.

References

[1] S. Wasserman and K. Faust, Social network analysis: Methods and applications. Cambridge Univ Pr, 1994.

[2] A. Bavelas, "A mathematical model of Group Structure," Human Organizations, vol. 7, pp. 16–30, 1948.

[3] L. C. Freeman, "Centrality in social networks: Conceptual clarification," Social Networks, vol. 1, no. 3, pp. 215–239.

[4] G. Nomikos, P. Pantazopoulos, M. Karaliopoulos, and I. Stavrakakis, "The multiple instances of node centrality and their implications on the vulnerability of ISP networks," Tech. Rep., 2013. [Online]. Available: *http://arxiv.org/abs/1312.4707*

[5] P. Pantazopoulos, M. Karaliopoulos, and I. Stavrakakis, "On the local approximations of node centrality in Internet router-level topologies," in 7th IFIP IWSOS, Palma de Mallorca, Spain, 2013.

[6] S. Borgatti and M. Everett, "A Graph-theoretic perspective on centrality," Social Networks, vol. 28, no. 4, pp. 466–484, Oct. 2006.

[7] N. T. Spring et al., "Measuring ISP topologies with rocketfuel." IEEE/ACM Trans. Netw., vol. 12, no. 1, pp. 2–16, 2004.

[8] The CAIDA UCSD Internet Topology Data Kit ITDK2011-10. [Online]. Available: *http://www.caida.org/data/active/internet-topology-data-kit/*

[9] J.-J. Pansiot et al., "Extracting intra-domain topology from mrinfo probing," in Proc. PAM, Zurich, Switzerland, April 2010.

[10] S. Knight et al., "The internet topology zoo." IEEE Journal on Selected Areas in Communications, vol. 29, no. 9, pp. 1765–1775, 2011.

[11] C.-Y. Lee, "Correlations among centrality measures in complex networks." [Online]. Available: *http://arxiv.org/abs/physics/0605220*

[12] A. Vázquez et al., "Large-scale topological and dynamical properties of the internet," Phys. Rev. E, vol. 65, no. 6, p. 066130, Jun 2002.

[13] V. Grolmusz, "A note on the pagerank of undirected graphs," May 2012. [Online]. Available: *http://arxiv.org/abs/1205.1960*

[14] E. Yan and Y. Ding, “Applying centrality measures to impact analysis: A coauthorship network analysis,” J. Am. Soc. Inf. Sci. Technol., vol. 60, no. 10, pp. 2107–2118, Oct. 2009.

[15] P. Holme, B. J. Kim, C. N. Yoon, and S. K. Han, “Attack vulnerability of complex networks,” Phys. Rev. E, vol. 65, no. 5, May 2002.

[16] R. Albert, H. Jeong, and A.-L. Barabasi, “Error and attack tolerance of complex networks,” Nature, vol. 406, no. 6794, pp. 378–382, Jul. 2000.

This thesis has been accepted for presentation at the 26th International Teletraffic Congress (ITC 2014) in Sweden on 9-11 September with the title “Comparative Assessment of Centrality Indices and Implications on the Vulnerability of ISP Networks”.

Βασίλειος Α. Τσιρώνης

grad0993@di.uoa.gr

Μελέτη Μηχανισμού Διασφάλισης Συνέπειας Εξυπηρετητών Κρυφής Μνήμης Παγκόσμιου Ιστού, με Εφαρμογή της Θεωρίας Βέλτιστης Παύσης

Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, Τμήμα Πληροφορικής και Τηλεπικοινωνιών
Πανεπιστημιούπολη, Ιλίσια, 15784, Αθήνα, Ελλάς

Περίληψη

Αντικείμενο της παρούσας εργασίας είναι η σχεδίαση και ανάπτυξη ενός μηχανισμού κρυφής μνήμης (cache) για τον Παγκόσμιο Ιστό (Web), ο οποίος θα βασίζεται στη θεωρία της Βέλτιστης Παύσης (Optimal Stopping). Πρόκειται για έναν μηχανισμό που θα αφορά κυρίως τους εξυπηρετητές κρυφής μνήμης (cache servers) και θα στοχεύει στη βελτιστοποίηση της επίλυσης του προβλήματος της συνέπειας της κρυφής μνήμης (cache consistency) σε αυτούς, μοντελοποιώντας το ως ένα πρόβλημα βέλτιστης επιλογής (best choice).

Λέξεις κλειδιά: Κρυφή μνήμη, Συνέπεια, Παγκόσμιος Ιστός, Εξυπηρετητές κρυφής μνήμης, Βέλτιστη παύση.

Επιβλέπων

Ευστάθιος Χατζηευθυμιάδης, Αναπληρωτής Καθηγητής

1. Εισαγωγή

Το Web είναι ένα πληροφοριακό σύστημα χτισμένο πάνω στο Διαδίκτυο (Internet), το οποίο παρέχει προσπέλαση σε διασυνδεδεμένα έγγραφα υπερκειμένου (hypertext) κάνοντας χρήση του HTTP πρωτοκόλλου. Η λειτουργία του είναι απλή: όποτε ένας web χρήστης[1] ζητά ένα αντικείμενο, μια HTTP αίτηση (request) αποστέλλεται προς ένα web εξυπηρετητή (server), ο οποίος αναλαμβάνει να στείλει την κατάλληλη HTTP απάντηση (response) [11].

Η διάδοση του Web στις μέρες μας είναι ραγδαία [1]. Η εξάπλωση του εκφράζεται μέσα από 2 τάσεις: α) την τεράστια αύξηση των χρηστών του Internet, β) την ολοένα και μεγαλύτερη ανάπτυξη τεχνολογιών γύρω από αυτό. Το πρόβλημα που άμεσα τίθεται λόγω της τρέχουσας πραγματικότητας, αφορά τη διατήρηση της επεκτασιμότητας (scalability) του Web. Δηλαδή, της ικανότητας του να διαχειρίζεται τον ολοένα και αυξανόμενο φόρτο εργασίας με επιδέξιο τρόπο [2,3].

Μια λύση στο παραπάνω ζήτημα αποτελούν οι web caches (browser caches, cache servers), οι οποίες είναι στοιχεία που βρίσκονται γεωγραφικά εγγύτερα στους web χρήστες και στόχο έχουν την άμεση εξυπηρέτηση των αιτήσεων τους, επιτυγχάνοντας με τον τρόπο αυτό μείωση στην καθυστέρηση απόκρισης (latency). Το πρόβλημα που προκύπτει από τη λύση αυτή ονομάζεται cache consistency και η περίπτωση του είναι ιδιαίτερη λόγω της κατανεμημένης φύσης του Web, αλλά και του μεγάλου όγκου του. Συγκεκριμένα, το πρόβλημα συνίσταται στην τροποποίηση ενός αντικειμένου στον εξυπηρετητή προέλευσης (origin server) και στη μη έγκαιρη ενημέρωση του αντιγράφου στην cache [4].

Πολλοί μηχανισμοί έχουν προταθεί και αναπτυχθεί για την επίλυση αυτού του προβλήματος. Γνωστότερος όλων είναι ο Adaptive TTL (ATTL), ο οποίος προσπαθεί μέσω εκχώρησης κατάλληλα διαμορφωμένων χρόνων ζωής (Time to Live, TTL) να διασφαλίσει μια ασθενή συνέπεια μεταξύ ενός αντικειμένου και των αντιγράφων του στις web caches. Στα πλαίσια αυτής της εργασίας μελετάται η δυνατότητα υλοποίησης ενός μηχανισμού που θα μοντελοποιεί το πρόβλημα της συνέπειας ως ένα πρόβλημα της θεωρίας Βέλτιστης Παύσης, με στόχο την αποδοτικότερη αντιμετώπιση του. Παρόμοια προσπάθεια γίνεται στην εργασία [10], όπου παρουσιάζεται ένας ασύγχρονος μηχανισμός συνέπειας βασισμένος

1. *Ο web χρήστης σε αυτή περίπτωση λειτουργεί ως client μιας και το Web έχει υιοθετήσει μια client - server αρχιτεκτονική, την οποία το HTTP αξιοποιεί μέσω του request - response μοντέλου.*

στον Odds αλγόριθμο. Στη παρούσα εργασία, όμως, το πρόβλημα μελετάται ευρύτερα, καθώς εξετάζεται η απόδοση ενός τέτοιου μηχανισμού σε συνάρτηση με συγκεκριμένα χαρακτηριστικά της μοντελοποίησης. Για το λόγο αυτό υλοποιείται ένας προσομοιωτής, ο Vproxy, μέσω του οποίου γίνεται προσπάθεια τεκμηρίωσης και αποτίμησης του προτεινόμενου μηχανισμού.

Το υπόλοιπο της εργασίας δομείται σε πέντε παραγράφους. Στην 2η παράγραφο γίνεται μια σύντομη εισαγωγή στον κόσμο του web caching και τους cache consistency[2] μηχανισμούς. Στην 3η παράγραφο περιγράφονται οι βασικές αρχές της θεωρίας Βέλτιστης Παύσης, απαραίτητες για την κατανόηση της 4ης παραγράφου όπου παρουσιάζεται η μοντελοποίηση του προτεινόμενου μηχανισμού. Τέλος, στην 5η παράγραφο παρουσιάζονται τα αποτελέσματα της προσομοίωσης, με τα συμπεράσματα που διεξήχθησαν από αυτή να καταγράφονται στη τελευταία παράγραφο.

Πριν, όμως, ο αναγνώστης διαβάσει τις επόμενες παραγράφους είναι απαραίτητο να αποσαφηνιστεί το εξής: επειδή η εργασία αυτή επικεντρώνεται στη λειτουργία των cache servers, οι οποίοι αναφέρονται στη βιβλιογραφία και ως web caching proxies ή χάριν απλούστευσης proxies, από το σημείο αυτό και έπειτα θα γίνεται χρήση του όρου proxy αντί του όρου web cache που χρησιμοποιήθηκε μέχρι τώρα, καθώς επίσης όλη η ανάλυση θα αφορά πλέον τη λειτουργία αυτών.

2. Κρυφή Μνήμη στο Web

Η λογική των proxies είναι ίδια με αυτή των caches στα PCs. Αναλαμβάνουν να ικανοποιήσουν τις αιτήσεις ενός web χρήστη με στόχο να απαλύνουν το φόρτο εργασίας (workload) στους origin servers. Όταν ένας proxy λάβει μια αίτηση έχει δύο επιλογές: είτε να ικανοποιήσει την αίτηση στέλνοντας στο χρήστη την κατάλληλη απάντηση, εφόσον διαθέτει το αντίστοιχο αντίγραφο το οποίο θεωρεί έγκυρο (valid) είτε να προωθήσει την αίτηση προς τον origin server. Στην 1η περίπτωση ο χρήστης επωφελείται από ένα συμβάν που ονομάζεται γρήγορη ευστοχία (fast hit). Η 2η περίπτωση μπορεί να συμβεί όταν ο proxy δεν διαθέτει κάποιο αντίγραφο και άρα συμβαίνει υποχρεωτική αστοχία (compulsory miss) ή όταν ο proxy κρίνει πως το αντίγραφο που διαθέτει δεν είναι έγκυρο

2. *Στην βιβλιογραφία μπορεί να συναντηθεί και με τον όρο cache coherence.*

και ακολούθως πρέπει να το επικυρώσει επικοινωνώντας με τον origin server. Εάν η επικύρωση επιβεβαιώσει την μη εγκυρότητα του αντιγράφου, τότε συμβαίνει αστοχία συνέπειας (consistency miss) και άρα πρέπει να ενημερωθεί ο proxy βάσει καινούριου αντιγράφου, ενώ εάν το αντικείμενο δεν έχει αλλάξει στον origin server το συμβάν ονομάζεται αργή ευστοχία (slow hit) [4].

Από τα παραπάνω γίνεται αντιληπτό ότι καλύτερη περίπτωση αποτελεί το fast hit, καθώς μέσω αυτού επιτυγχάνεται αποσυμφόρηση δικτύου, μείωση workload στους origin servers, μείωση latency στους χρήστες και εξοικονόμηση εύρους ζώνης (bandwidth), συμβάλλοντας έτσι στη ποθητή επεκτασιμότητα του Web. Από την άλλη, σε ένα fast hit ελλοχεύει πάντα ο κίνδυνος παράδοσης έωλου αντικειμένου (stale delivery), που σημαίνει την ικανοποίηση, από πλευράς ενός proxy, μιας HTTP αίτησης κάποιου χρήστη μέσω ενός αντιγράφου που θεωρεί έγκυρο, ενώ στην πραγματικότητα το αντικείμενο έχει τροποποιηθεί στον origin server.

Μια λύση στο πρόβλημα του stale delivery αποτελεί η χρήση ισχυρών μηχανισμών συνέπειας. Οι μηχανισμοί αυτοί επιβάλουν άμεση ενημέρωση των αντιγράφων ενός αντικειμένου όταν αυτό τροποποιηθεί στον origin server. Ένας γνωστός μηχανισμός της κατηγορίας αυτή είναι ο Client Validation [4]. Πιο γνωστοί, όμως, είναι οι Polling Every Time και Server Invalidation, οι οποίοι ανήκουν στην κατηγορία των μηχανισμών ακύρωσης (invalidation) και στους οποίους την ευθύνη για την ενημέρωση των αντιγράφων στους proxies αναλαμβάνει ο origin server [4]. Στην αντίπερα όχθη των ισχυρών μηχανισμών βρίσκονται οι μηχανισμοί επικύρωσης (validation), οι οποίοι επιβάλλουν ασθενή συνέπεια στα δεδομένα μιας cache. Γνωστοί μηχανισμοί της κατηγορίας αυτής είναι ο ATTL που έχει αναφερθεί πιο πάνω και θα αναλυθεί στη συνέχεια, καθώς και ο Piggyback Cache Validation (PCV), ο οποίος αποτελεί μια προσέγγιση ασύγχρονου (asynchronous) μηχανισμού [4]. Στους μηχανισμούς αυτούς την ευθύνη για την ενημέρωση των αντιγράφων την έχουν οι proxies και όχι οι servers, και αυτό το χαρακτηριστικό τους κάνει πιο ελκυστικούς στο Web, αφού από τη μια επιβαρύνουν λιγότερο το δίκτυο και από την άλλη το HTTP μπορεί να τους υποστηρίξει καλύτερα [11].

Σε αυτό το σημείο καθίσταται αναγκαίο να περιγραφεί αναλυτικότερα η λειτουργία του ATTL, καθώς είναι ο σημαντικότερος μηχανισμός από όλους τους παραπάνω. Ο μηχανισμός αυτός οφείλει την προέλευση του στο πρωτόκολλο Alex [5] και βασίζεται σε μια πολύ απλή λογική: «όσο περισσότερο έχει παραμείνει ένα αντικείμενο αμετάβλητο τόσο περισσότερο τείνει να παραμείνει αμετάβλητο στο μέλλον». Βάσει αυτού του μηχανισμού, λοιπόν, όταν ένας

proxy λαμβάνει ένα αντικείμενο, υπολογίζει το σχετικό του TTL ως ένα κλάσμα του χρόνου που έχει κυλήσει μεταξύ της τελευταίας τροποποίησης του και του χρόνου που απεστάλει από τον origin server [4]:

$$TTL = \min\{k \times (send\_time - last\_\mathrm{mod}ified), threshold\}$$

όπου το k είναι μια σταθερά με τυπικές τιμές 0.1 ή 0.2 και το $threshold$, ένα κατώφλι το οποίο χρησιμοποιείται για να εξασφαλιστεί ότι ακόμα και τα αντικείμενα τα οποία δεν έχουν τροποποιηθεί για μεγάλο χρονικό διάστημα θα ελεγχθούν. Στη συνέχεια, ικανοποιεί όλες τις αιτήσεις που αφορούν αυτό το αντικείμενο κατά τη διάρκεια του TTL, ενώ όταν αυτό λήξει στέλνει μια αίτηση επικύρωσης (validation request) ώστε να διασφαλίσει τη συνέπεια του.

3. Θεωρία Βέλτιστης Παύσης

Τα προβλήματα με τα οποία ασχολείται η θεωρία Βέλτιστης Παύσης ονομάζονται προβλήματα εύρεσης κανόνα παύσης (stopping rule) και ορίζονται από τα εξής δύο αντικείμενα [6]:

- μια ακολουθία τυχαίων μεταβλητών (τμ) $X_1, X_2, \ldots,$ των οποίων η κοινή κατανομή (joint distribution) θεωρείται γνωστή, και
- από μια ακολουθία πραγματικών συναρτήσεων ανταμοιβής (reward functions) $y_0, y_1(x_1), y_1(x_1, x_2), \ldots, y_\infty(x_1, x_2, \ldots)$.

Βάσει, λοιπόν, του παραπάνω ορισμού ο κανόνας παύσης μπορεί να διατυπωθεί απλούστερα ως εξής:

(1) Μπορείς να παρατηρήσεις τις $X_1, X_2, \ldots,$ για όσο διάστημα θες,

(2) Σε κάθε βήμα $n = 1, 2, \ldots,$ έχοντας παρατηρήσει $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, μπορείς είτε να σταματήσεις και να λάβεις ως ανταμοιβή την τιμή της συνάρτησης $y_n(x_1, x_2, \ldots, x_n)$, είτε να συνεχίσεις και να παρατηρήσεις την X_{n+1},

(3) Εάν επιλέξεις να σταματήσεις δίχως να κάνεις παρατηρήσεις, τότε θα λάβεις την σταθερή τιμή y_0,

(4) Εάν δεν σταματήσεις ποτέ, τότε θα λάβεις $y_\infty(x_1, x_2, \ldots)$.

Στόχος είναι η επιλογή εκείνου του χρόνου παύσης που θα μεγιστοποιήσει την αναμενόμενη ανταμοιβή ή θα ελαχιστοποιήσει το αναμενόμενο κόστος.

Σημαντικό παράδειγμα προβλήματος που επιλύεται με χρήση κανόνα παύσης είναι αυτό της πώλησης σπιτιού (house selling problem) [6], όπου προσφορές καταφθάνουν σε τυχαία χρονικά διαστήματα και στόχος είναι η πώληση στη μεγαλύτερη τιμή. Ο πωλητής δεν έχει καμία a priori γνώση για τις τιμές των προσφορών, παρά μόνο μια διαίσθηση πως αυτές είναι ανεξάρτητες μεταξύ τους και διέπονται από μια κοινή κατανομή. Σε κάθε προσφορά έχει 2 επιλογές: 1) αποδοχή και πέρας αγοραπωλησίας, 2) απόρριψη και αναμονή για την επόμενη προσφορά πληρώνοντας ένα σταθερό κόστος c. Το δίλλημα, λοιπόν, συνίσταται στην επιλογή εκείνης της προσφοράς που θα ναι μεν μεγαλύτερη από τις προηγούμενες, θα αντισταθμίζει δε τα αντίστοιχα κόστη παρατήρησης.

Ένα άλλο γνωστό πρόβλημα της θεωρίας αυτής είναι το κλασσικό πρόβλημα επιλογής γραμματέως (classical secretary problem, CSP). Το πρόβλημα αυτό ορίζει μια ολόκληρη κατηγορία προβλημάτων τα οποία ονομάζονται προβλήματα βέλτιστης επιλογής, διότι η μόνη αποδεκτή λύση σε αυτά είναι η καλύτερη. Στο CSP για παράδειγμα, στόχος είναι η επιλογή του καλύτερου μέσα από ένα πλήθος n αιτούντων. Αποδεικνύεται ότι ο κανόνας παύσης που δίνει τη λύση στο CSP είναι ο ακόλουθος [6]:

$$N_1 = \min\{ j \geq 1 : \sum_{k=j+1}^{n} \frac{1}{k-1} \leq 1 \}$$

Ο παραπάνω κανόνας ονομάζεται και κανόνας κατωφλίου (threshold rule), διότι αυτό που ουσιαστικά κάνει είναι να προσπερνά τους πρώτους j αιτούντες (όσο δηλαδή ορίζει το κατώφλι) και στη συνέχεια να επιλέγει τον πρώτο υποψήφιο (candidate)[3] που θα συναντήσει.

Εάν το n είναι πολύ μεγάλο τότε είναι προσεγγιστικά βέλτιστο να προσπεράσει ο κανόνας παύσης το $e^{-1}=36.8\%$ των αιτούντων και στη συνέχεια να επιλέξει τον πρώτο υποψήφιο που θα συναντήσει, με πιθανότητας νίκης e^{-1} [6].

Δύο κανόνες που ορίζουν ένα πιο γενικό πλαίσιο επίλυσης προβλημάτων όπως το προηγούμενο και οι οποίοι θα αναλυθούν στην επόμενη παράγραφο, καθώς θα

3. *Ως υποψήφιος ορίζεται εκείνος ο αιτών που είναι καλύτερος από όλους τους αιτούντες που έχουν εξεταστεί σε ένα δεδομένο χρονικό διάστημα.*

χρησιμοποιηθούν στη μοντελοποίηση του προτεινόμενου μηχανισμού, είναι ο 1/e κανόνας [7], καθώς και ο Odds αλγόριθμος [8].

4. Μοντελοποίηση Μηχανισμού

Ο μηχανισμός συνέπειας που θα προταθεί και θα περιγραφεί στη συνέχεια του κειμένου ονομάζεται Vproxy. Ουσιαστικά δανείζεται το όνομα του από τον προσομοιωτή που αναπτύχθηκε για τον πειραματικό έλεγχο της συμπεριφοράς και απόδοσης του. Από αυτό το σημείο, λοιπόν, με τον όρο Vproxy θα εννοείται ο ίδιος ο μηχανισμός.

Ο Vproxy είναι ένας μηχανισμός ασθενής, επικύρωσης (validation) και σύγχρονος. Έχει, δηλαδή, όλα εκείνα τα χαρακτηριστικά που έχει και ο ATTL. Στόχος του, μάλιστα, είναι η περαιτέρω βελτίωση του ATTL με ταυτόχρονη αξιοποίηση των πλεονεκτημάτων του. Για να το καταφέρει αυτό προσπαθεί να βελτιώσει τη διαχείριση του tradeoff που είναι εγγενές σε κάθε μηχανισμό συνέπειας και το οποίο συντίθεται από τη διλημματική επιλογή ανάμεσα στο κόστος επικοινωνίας και το βαθμό συνέπειας. Πιο συγκριμένα, ισχυρότερος βαθμός συνέπειας σημαίνει μεγαλύτερο κόστος επικοινωνίας, ενώ μικρότερο κόστος επικοινωνίας συνεπάγεται ασθενέστερο βαθμό συνέπειας. Άρα, το πρόβλημα πλέον συνίσταται στην εύρεση εκείνου του μηχανισμού που θα επιτυγχάνει το μεγαλύτερο δυνατό βαθμό συνέπειας με το μικρότερο δυνατό κόστος επικοινωνίας. Ένας τέτοιος μηχανισμός είναι και ο ATTL.

Την παραπάνω κατάσταση, ο Vproxy την αντιμετωπίζει μοντελοποιώντας τη διαχείριση του tradeoff ως ένα πρόβλημα βέλτιστης παύσης. Για το λόγο αυτό, προτείνει την έννοια του cacheability για ένα αντικείμενο που βρίσκεται σε μια cache, ως το μέγεθος εκείνο που καθορίζει την δυνατότητα χρήσης αυτού του αντικειμένου προς ικανοποίηση μιας αίτησης. Στη συνέχεια, υποθέτει πως εάν ελέγξει τη συνέπεια ενός αντικειμένου μιας cache ενός proxy την ακριβή εκείνη στιγμή που θα παρατηρηθεί η μέγιστη τιμή στο cacheability του, ύστερα από την οποία αυτό (το cacheability) θα φθίνει και άρα η πιθανότητα stale delivery θα αυξάνει, τότε θα επιτευχθεί η βέλτιστη αναλογία μεταξύ slow hits (κόστος επικοινωνίας) και stale deliveries (βαθμός συνέπειας). Άρα μιλάμε για ένα πρόβλημα βέλτιστης επιλογής, διότι η μόνη αποδεκτή λύση είναι ο έλεγχος της συνέπειας στη πλέον cacheable κατάσταση ενός αντικειμένου.

Επειδή ο Vproxy είναι σύγχρονος μηχανισμός, μια παρατήρηση λαμβάνεται

κατά τη διάρκεια μιας αίτησης. Με άλλα λόγια, κάθε αίτηση συνιστά πλέον και μια παρατήρηση. Κατά τη διάρκεια, λοιπόν, μιας αίτησης προς ένα αντικείμενο αποτιμάται μια μετρική η οποία αποτελεί την ποσοτική εκτίμηση, βάσει συγκεκριμένων χαρακτηριστικών, της διαίσθησης που σχετίζεται με το cacheability του. Άρα, η τιμή της μετρικής για την τρέχουσα αίτηση αποτελεί την τρέχουσα παρατήρηση. Εάν η τιμή αυτή είναι η μεγαλύτερη που έχει παρατηρηθεί για το συγκεκριμένο αντικείμενο ελέγχεται ο κανόνας παύσης, ώστε να αποφασιστεί εάν είναι ταυτόχρονα η μέγιστη τιμή, ούτως ώστε να γίνει έλεγχος συνέπειας. Διαφορετικά, η αίτηση ικανοποιείται από τον proxy.

Συνεπώς, το συγκεκριμένο πρόβλημα βέλτιστης παύσης για κάθε αντικείμενο μιας cache ενός proxy ορίζεται ως εξής:

- από την ακολουθία των τιμών $U_1, U_2,...$ μιας μετρικής U
- από την ακολουθία των συναρτήσεων ανταμοιβής που ορίζονται ως εξής:

$$y(u_i) = \begin{cases} 0, & u_i \le M_i \\ P(U_{i+1} \le u_i, U_{i+2} \le u_i, ...), & u_i > M_i \end{cases}$$, όπου $M_i = max\ \{U_0, ... U_{i-1}\}$

Στόχος, λοιπόν, του κανόνα παύσης είναι η επιλογή του χρόνου εκείνου που μεγιστοποιεί την πιθανότητα η παρατηρηθείσα μέγιστη τιμή να είναι όντως η μέγιστη.

4.1. Μετρικές

Οι μετρικές που χρησιμοποιούνται στη παρούσα μοντελοποίηση είναι οι u_3 και u_4 και αποτελούν τη σύνθεση των επιμέρους μετρικών: δημοτικότητα αντικειμένου (object popularity, op), ποσοστό ευστοχίας ιστοτόπου (site hit ratio, shr), αναλογία χρόνου της αίτησης (request time ration, rtr) και μεταβλητότητα αντικειμένου (object mutability, om). Η u_4 συντίθεται με βάση τις 3 πρώτες μετρικές, ενώ η u_4 επεκτείνει την u_4 προσθέτοντας την om. Οι παραπάνω μετρικές ορίζονται ως εξής:

$u_3 = w_1 \times shr + w_2 \times op + w_3 \times rtr$ και $u_4 = u_3 + w_4 \times om$, όπου

$$op = \frac{object\ requests}{object's site\ requests}, \quad shr = \frac{site\ hits}{site\ total\ requests}$$

$$rtr = \min\{\frac{t_{current} - t_{load}}{t_{\exp ires} - t_{current}}, 1\}, \quad om = \frac{object \bmod ifications}{object\ requests}$$

Τα βάρη w_i στις παραπάνω μετρικές είναι ισοσκελισμένα στις τιμές 1/3 για την u_3 και 1/4 για την u_4.

4.2. Κανόνες Παύσης

Ο 1ος κανόνας παύσης που υλοποιείται από τον Vproxy είναι 1/e ο οποίος προτάθηκε από τον F. Thomas Bruss το 1984, με σκοπό την ενοποίηση των προβλημάτων βέλτιστης επιλογής κάτω από την πλήρη άγνοια του αριθμού των παρατηρήσεων, αλλά και της κατανομής αυτού [7]. Η λογική του είναι πολύ απλή:

Έστω ότι η χρονική στιγμή άφιξης κάθε αιτούντα είναι iid[4] με pdf f στο διάστημα [0,T] και cdf F τέτοια ώστε να ισχύει:

$$F(t) = \int_0^t f(s)ds, \quad 0 \le t \le T,$$

και έστω χρόνος τ τέτοιος ώστε να ισχύει $F(\tau) = 1/e$, τότε είναι βέλτιστο να περιμένεις μέχρι την χρονική στιγμή τ και στη συνέχεια να επιλέξεις τον πρώτο υποψήφιο που θα συναντήσεις [7].

Εάν υποτεθεί ότι οι χρόνοι άφιξης των αιτούντων είναι ομοιόμορφα idd στο [0,T] τότε ο 1/e κανόνας θα επιλέγει πάντα τον 1ο υποψήφιο με χρόνο άφιξης $t \ge e^{-1}T$.

Ο 2ος κανόνας παύσης προκύπτει από το Odds Theorem [8], όπως αυτό προτάθηκε από τον F. Thomas Bruss σε μια προσπάθεια να δοθεί ένα γενικότερο πλαίσιο επίλυσης προβλημάτων βέλτιστης επιλογής, όπου οι παρατηρήσεις μπορούν να πάρουν τιμή 1 (επιτυχία) η τιμή 0 (αποτυχία), και δεν έχουν Markov δομή. Το Odds θεώρημα, λοιπόν, λέει [8]:

Έστω ότι η ακολουθία $I_1, I_2, \ldots, I_n$ είναι μια ακολουθία ανεξάρτητων

4. *Identically independently distributed*

δείκτριων συναρτήσεων με $p_j = E(I_j)$ και $q_j = 1 - p_j$, $r_j = p_j / q_j$, τότε υπάρχει ένας βέλτιστος κανόνας για παύση στην τελευταία επιτυχία και είναι η παύση στην 1η επιτυχία $k(I_k=1)$, για την οποία ισχύει $K \geq s$, όπου:

$$s = \sup\{1, \sup\{1 \leq k \leq s : \sum_{j=k}^{n} r_j \geq 1\}\} \text{ , με } \sup\{\} = -\infty$$

Η πιθανότητα νίκης δίνεται από τη σχέση: $V(n) = \prod_{j=s}^{n} q_j \sum_{j=s}^{n} r_j$ [8]

Το παραπάνω θεώρημα, μας δίνει τον Odds κανόνα παύσης ο οποίος συνοπτικά μας λέει να σταματήσουμε όταν το άθροισμα φτάσει ή ξεπεράσει την τιμή 1 ή όταν ο κανόνας φτάσει στην τιμή $k=1$.

Στο πρόβλημα μας, όμως, ο χρόνος είναι συνεχής και οι παρατηρήσεις γίνονται σε τυχαία χρονικά διαστήματα, όπου τον ορίζοντα τον ορίζει το TTL του αντικειμένου. Οι δε αφίξεις των αιτήσεων προς ένα αντικείμενο ακολουθούν ομοιογενή Poisson διεργασία [12]. Επιπλέον, το συγκεκριμένο πρόβλημα βέλτιστης επιλογής ανήκει στην κατηγορία των full-information προβλημάτων, μιας και οι τιμές των τμ U είναι πραγματικές και iid στο [0,1] . Για όλους τους παραπάνω λόγους ο Odds που χρησιμοποιείται από τον Vproxy ορίζεται ως εξής:

$$\lambda \times (1-u) \times t \leq 0.80435...,$$

όπου t είναι ο υπολειπόμενος χρόνος, δηλαδή ο χρόνος που απομένει μέχρι τη λήξη του TTL, λ είναι ο ρυθμός έλευσης των αιτήσεων (ένταση των Poisson αφίξεων) και u η μέγιστη παρατηρηθείσα τιμή της μετρικής.

Στα πλαίσια της εργασία αυτής αναπτύχθηκε μια απόδειξη του παραπάνω τύπου, κάνοντας χρήση του Odds θεωρήματος. Η παραπάνω ανισότητα συμφωνεί απόλυτα με τη λύση που έχουν δώσει οι Sakaguchi [13] και Bodjecki [9].

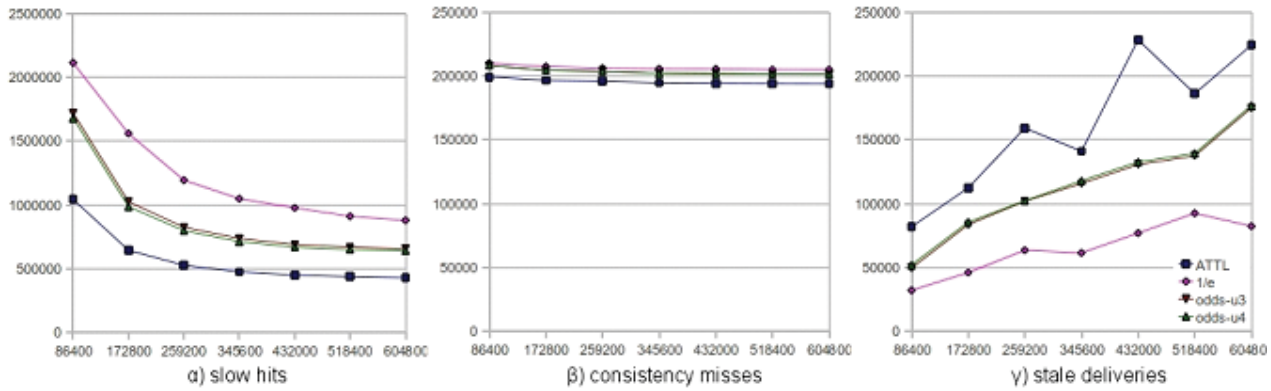
5. Προσομοίωση

Για την προσομοίωση του Vproxy ακολουθήθηκε η μεθοδολογία εκείνη που βασίζεται στη χρήση trace αρχείων. Συγκεκριμένα, χρησιμοποιήθηκαν Web traces της DEC τα οποία αφορούσαν μια περίοδο 3.5 εβδομάδων του 1996 και

περιλαμβάνουν 24.477.674 αναφορές.

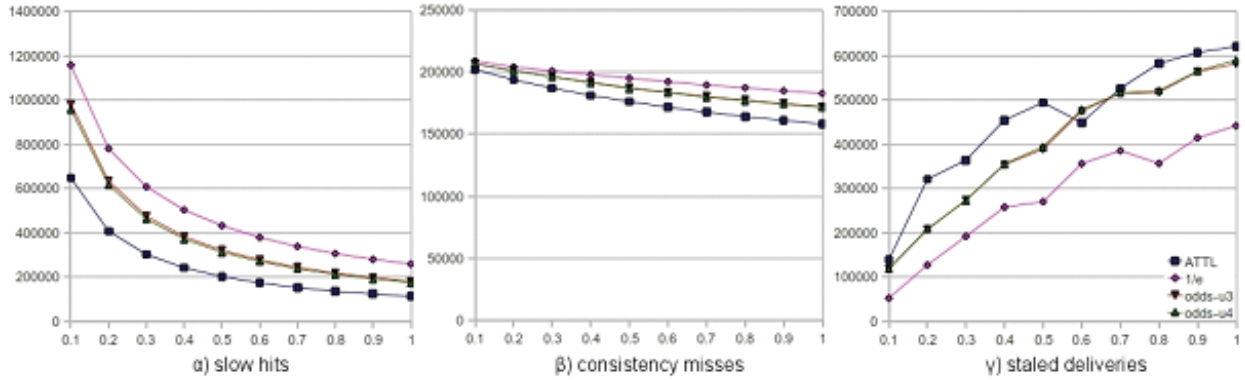
Έγιναν διάφορες προσομοιώσεις για τον έλεγχο της απόδοσης του Vproxy σε σύγκριση με τον ATTL. Η απόδοση του μετρήθηκε με βάσει τα slow hits, consistency misses και stale deliveries. Επιπλέον, μελετήθηκε η επίδραση διαφορών παραμέτρων στην απόδοση του, όπως: μετρική, κανόνας παύσης, πολιτική απόδοσης TTL, παράγοντας λ[5]. Στην πρώτη σειρά προσομοιώσεων, ο Vproxy χρησιμοποίησε τον ATTL ως υπόστρωμα, δηλαδή ως πολιτική απόδοσης TTL.

Από τα αποτελέσματα της πρώτης σειράς προσομοιώσεων προκύπτουν ενδιαφέροντα συμπεράσματα (βλέπε σχ.1 & σχ.2). Πρώτον είναι εμφανής η βελτιωμένη απόδοση του Vproxy σε σχέση με τον ATTL. Για κάθε κανόνα παύσης και κάθε μετρική πετυχαίνει καλύτερη απόδοση όσον αφορά τα stale deliveries και τα consistency misses. Μάλιστα, την καλύτερη επίδοση την έχει για τον $1/e$ την οποία όμως πληρώνει με τεράστια αύξηση στον αριθμό των slow hits, περίπου 142%. Από την άλλη, ο Vproxy με Odds κανόνα πετυχαίνει εξίσου μεγάλη μείωση στα stale deliveries με αντίστοιχο κόστος την αύξηση των slow hits μόνο κατά 50% περίπου. Για παράδειγμα, ο Vproxy με Odds κανόνα και u_3 μετρική πετυχαίνει μείωση στα stale deliveries κατά 43% με αντίστοιχη αύξηση 3.83% των consistency misses και 53.47% των slow hits.



Σχήμα 1: Μέτρηση κόστους για διάφορες τιμές threshold

5. *Αυτός αφορά κυρίως την απόδοση του Odds κανόνα και συνεπώς την απόδοση του Vproxy για αυτό τον κανόνα.*



Σχήμα 2: Μέτρηση κόστους για διάφορες τιμές k

Στις επόμενες σειρές προσομοιώσεων εξετάζεται η επίδραση του λ στην απόδοση του Vproxy για Odds κανόνα, καθώς και της πολιτικής απόδοσης TTL. Αυτό που γίνεται έντονα σαφές είναι πως και τα δύο μεγέθη επιδρούν καθοριστικά, καθώς είτε μια εσφαλμένη εκτίμηση του λ είτε μια αφελής πολιτική εκχώρησης TTL μπορούν να οδηγήσουν σε σοβαρή υποβάθμιση της απόδοσης του.

6. Συμπεράσματα

Από την προσομοίωση προέκυψε ότι ο Vproxy παρουσιάζει αξιόλογη απόδοση όταν χρησιμοποιεί ως υπόστρωμα τον ATTL. Μάλιστα, με χρήση Odds κανόνα παρουσίασε πιο ισορροπημένη συμπεριφορά, αφού μείωσε δραματικά τα stale deliveries με μικρό κόστος σε slow hits. Το δε κόστος σε slow hits είναι αναπόφευκτο, μιας και είναι σύμφυτο με το tradeoff που περιγράφηκε στην 4η παράγραφο. Δεν παρουσίασε, όμως, την ίδια απόδοση όταν χρησιμοποιήθηκε μια απλοϊκή πολιτική απόδοσης σταθερών TTL σε όλα τα αντικείμενα. Άρα, για να είναι αποδοτικός ο Vproxy χρειάζεται μια ορθολογική πολιτική εκχώρησης TTL μιας και αυτά επηρεάζουν άμεσα τους κανόνες παύσης.

Μια άλλη ενδιαφέρουσα παρατήρηση αφορά τις μετρικές. Τα αποτελέσματα έδειξαν μικρή διαφορά στην απόδοση του Odds για τις 2 μετρικές. Αυτό σημαίνει ότι ο ορισμός μια μετρικής χρειάζεται επιπλέον μελέτη (π.χ. μελέτη της κατανομής). Σε παρόμοιο συμπέρασμα μας οδήγησε η μελέτη του λ. Θεωρήθηκε ότι η διαδικασία αφίξεων των αιτήσεων είναι ομοιογενής Poisson και πως το λ είναι η ένταση αυτής. Ίσως είναι αναγκαία η μελέτη του προβλήματος για ανομοιογενή διαδικασία Poisson.

Τέλος, επισημάνθηκε το φαινόμενο stale deliveries burst και η μεγάλη επίδραση του στην απόδοση ενός μηχανισμού συνέπειας. Το φαινόμενο αυτό αφορά τα δημοφιλή αντικείμενα και παρατηρείται όποτε ένα από αυτά τροποποιείται στον origin server, αλλά ο μηχανισμός συνέπειας εξακολουθεί να ικανοποιεί τις αιτήσεις προς αυτό χρησιμοποιώντας παρωχημένο αντίγραφο μιας cache. Σε αυτή τη περίπτωση, το σύστημα θα επιβαρυνθεί από ένα μεγάλο αριθμό stale deliveries. Το φαινόμενο αυτό αντιμετωπίζεται στον Vproxy για Odds κανόνα με χρήση κατάλληλου ευρετικού (heuristic), παραμένει όμως έντονο όταν εμφανίζεται σε ένα δημοφιλές αντικείμενο με μικρό βαθμό μεταβλητότητας.

Αναφορές

[1] "Internet World Stats", 30 Jun. 2012; *http://www.internetworldstats.com/stats.htm* [Προσπελάστηκε 02/2/13]

[2] Andre B. Bondi, "Characteristics of scalability and their impact on performance," Proc. 2nd Int'l workshop Software and performance (WOSP '00), ACM, 2000, 195-203.

[3] Mark D. Hill, "What is scalability?," SIGARCH Comp. Archit. News 18, Dec 1990, 18-21

[4] Michael Rabinovich and Oliver Spatschek, Web Caching and Replication, Addison-Wesley Longman Publishing Co., 2002.

[5] V. Cate, "Alex - a global filesystem," Proc. USENIX File Systems Workshop, 1992, 1-12

[6] Thomas S. Ferguson, "Optimal Stopping and Applications", UCLA, *http://www.math.ucla.edu/tom/Stopping/Contents.html.* [Προσπελάστηκε 02/2/13]

[7] F. T. Bruss, "A Unified Approach to a Class of Best Choice Problems with an Un known Number of Options," Ann. Probab, vol. 12, no. 3, 1984, 882-889

[8] F. T. Bruss, "Sum the odds to one and stop," Ann. Probab, vol. 28, no. 3, 2000, 1384-1391

[9] T. Bojdecki, "On optimal stopping of a sequence of independent random variables-

probability maximizing approach," Stochastic Processes and their Applications, vol. 6, no. 2, January 1978,153-163

[10] M. Spanoudakis, D. Lorentzos, C. Anagnostopoulos, S. Hadjiefthymiades, "Use of the Optimal Stopping Theory for Improving Cache Consistency", 13th Int'l Conf. on Web Information System Engineering (WISE 2012), Paphos, Cyprus, November, 2012

[11] R. Fielding et al., Hypertext Transfer Protocol - HTTP/1.1, IETF RFC 2616, June1999; *http://tools.ietf.org/rfc/rfc2616.txt*

[12] M. Arlitt, R. Friedrich and T. Jin, Workload characterization of a Web proxy in a cable modem environment, tech. report HPL-1999-48, Hewlett Packard Labs, 1999

[13] M. Sakaguchi, "Optimal stopping problems for randomly arriving offers," Mathematicae Japonicae, vol. 21, 1976, 201-217