Exploratory Data Analytics

Εργασία του μαθήματος Εφαρμογές Αναλυτικής Μεγάλων Δεδομένων

Αμερικάνος Πάρις – Πανεπιστήμιο Πειραιώς

Γενικά

Στην εργασία αυτή γίνεται εισαγωγή στη Διερευνητική Αναλυτική Δεδομένων με σκοπό την ανάλυση ενός χρονικά μεταβαλλόμενου συνόλου δεδομένων και τη βραχυπρόθεσμη πρόβλεψη των τιμών του. Γίνεται σύντομη αναφορά στη θεωρία ανάλυσης χρονοσειρών και διαφόρων μεθόδων αυτής, αναπτύσσεται η διαδικασία ανάγνωσης του αρχείου των δεδομένων, η προεπεξεργασία τους, η ανάλυσή τους και διερευνάται η πρόβλεψη των δυνατών τιμών, με ταυτόχρονη αποτίμηση της ποιότητας των προβλέψεων. Τέλος, εξετάζονται οι αλλαγές στη μεθοδολογία επίλυσης του προβλήματος για αρχεία εισόδου πολλαπλάσιου μεγέθους.

Σκοπός

Σκοπός της παρούσας εργασίας είναι η εξοικείωση με μεθόδους Διερευνητικής Αναλυτικής Δεδομένων, όπως σχηματισμός υποθέσεων, επιλογή κατάλληλων στατιστικών εργαλείων και αλγορίθμων, οπτικοποίηση δεδομένων εισόδου/εξόδου και εκτίμηση προβλέψεων. Η ανάλυση θα υλοποιηθεί πάνω σε μια τακτικά ενημερωμένη χρονοσειρά που αντιπροσωπεύει το πλήθος ετήσιων επιστημονικών δημοσιεύσεων από το 1936, με στόχο την πρόβλεψη του πλήθους δημοσιεύσεων για το ερχόμενο έτος. Ακόμη, θα συγκριθούν ως προς την ακρίβειά τους διάφορες μέθοδοι πρόβλεψης για την επιλογή της βέλτιστης, και θα παρατηρηθούν προβλήματα που δημιουργούνται από διαφορετικού μεγέθους σύνολα δεδομένων.

Θεωρία

Διερευνητική Αναλυτική Δεδομένων

Exploratory Data Analysis (ΔΑΔ) είναι μια προσέγγιση στην Ανάλυση Δεδομένων που αποφεύγει τις συνήθεις υποθέσεις σχετικά με το μοντέλο στο οποίο πρέπει να υπακούει ένα σύνολο δεδομένων, αλλά, στηριζόμενη κυρίως στις οπτικές και ποσοτικές μεθόδους της, επιτρέπει στα ίδια τα δεδομένα να αποκαλύψουν την αφανή δομή τους (Tukey, 1977) (Hartwig, 1979). Οι κύριες τεχνικές που χρησιμοποιούνται κατά τη ΔΑΔ είναι η οπτικοποίηση των δεδομένων (ιστόγραμμα, συνάρτηση πιθανότητας, κ.ά), η προεπεξεργασία τους (outliers, ελλιπείς ή λανθασμένες τιμές, κ.ά), η σχεδίαση απλών στατιστικών μέτρων (μέση τιμή, διασπορά, κλπ) και τέλος, η τοποθέτηση και ο μετασχηματισμός αυτών των διαγραμμάτων με τέτοιο τρόπο ώστε να μεγιστοποιείται η ανθρώπινη δυνατότητα αναγνώρισης μοτίβων και χαρακτηριστικών (NIST/SEMATECH, n.d.). Η ΔΑΔ δε χαρακτηρίζεται από αυστηρή μεθοδολογία για την εκτέλεσή της και βοηθά στη μακροσκοπική μελέτη των δεδομένων, παρέχοντας μια ευρεία οπτική σε μοτίβα, ανωμαλίες και χαρακτηριστικά των δεδομένων που θα οδηγήσουν στο σχηματισμό χρήσιμων υποθέσεων και απαντήσεων.

Χρονοσειρές

Χρονοσειρά είναι μια ακολουθία (περιοδική ή μη) δεδομένων διακριτού χρόνου τοποθετημένα σε χρονολογική σειρά. Συνήθως σχεδιάζεται σε διάγραμμα γραμμής και απαντάται σε όλους τους τομείς των επιστημών που περιλαμβάνουν με περιοδικές μετρήσεις. Η ανάλυση χρονοσειρών ασχολείται με την εξαγωγή χρήσιμων στατιστικών και χαρακτηριστικών από τα δεδομένα, ενώ στην πρόβλεψη χρονοσειρών γίνεται χρήση μοντέλων που μπορούν να προβλέψουν μελλοντικές τιμές των παρατηρούμενων μεγεθών βάσει παλαιότερων τιμών τους.

Ανάλυση και πρόβλεψη

Για την ανάλυση μιας χρονοσειράς υπάρχουν πολλές μέθοδοι διαθέσιμες, ανάλογα με τα χαρακτηριστικά της (περιοδικότητα, στατικότητα, γραμμικότητα, κά). Η περιγραφή μιας χρονοσειράς βάσει ενός μαθηματικού μοντέλου μπορεί να απλοποιήσει εξαιρετικά την ανάλυσή της και να παρέχει ακριβέστερη δυνατότητα πρόβλεψης, αποσυνθέτοντας (decomposition) αυτή σε 4 χαρακτηριστικά: τάση (μακροσκοπική συμπεριφορά), εποχικότητα (αυστηρή χρονική περίοδος), κυκλικότητα (μη περιοδικές διακυμάνσεις) και θόρυβος. Μια συνήθης και χρήσιμη μέθοδος είναι εύρεση καμπύλης ή συνάρτησης (curve fitting) η οποία προσαρμόζεται βέλτιστα στα δεδομένα και μπορεί να αντικαταστήσει ελλιπή δεδομένα ή να τονίσει τη συνδιακύμανση δυο μεταβλητών.

Στην πρόβλεψη μελλοντικών τιμών μια χρονοσειράς υπάρχουν επίσης πολλές μέθοδοι

διαφορετικής πολυπλοκότητας στηριζόμενες σε διαφορετικές παραμέτρους και αλγορίθμους. Προϋποτίθεται η στατικότητα της ακολουθίας για την εφαρμογή των προβλέψεων (σταθερή μέση τιμή και διακύμανση), διότι διαφορετικά το μοντέλο είναι σχεδόν τυχαίο (πχ τιμή μετοχής, καιρικές μετρήσεις). Συνήθεις μέθοδοι είναι απλή πρόβλεψη με βάση πρόσφατες μέσες τιμές (Singh, 2018), γραμμική τάση Holt (Kalekar, 2004), ARIMA με κυλιόμενο μέσο όρο και αυτοπαλινδρόμηση, κά (Chatfield, 2000).

Ανάλυση

Σύνολο δεδομένων

Το σύνολο δεδομένων που υποδείχθηκε για ανάλυση είναι το αρχείο DBLP μορφής XML μεγέθους 2.1GB που περιέχει βιβλιογραφικές αναφορές για όλες τις επιστημονικές δημοσιεύσεις πάνω στον κλάδο του Computer Science από το 1936 έως σήμερα, καταρτισμένο από το Πανεπιστήμιο του Trier (University of Trier, n.d.). Περιλαμβάνει εγγραφές για 4+εκ. δημοσιεύσεις από 2+εκ. συγγραφείς, είναι ελεύθερα προσβάσιμο και ανανεώνεται σε καθημερινή βάση. Η κάθε εγγραφή περιέχει τις πληροφορίες κάθε δημοσίευσης, συμπεριλαμβανομένου του έτους έκδοσης, βάσει του οποίου θα αναπτυχθεί μια χρονοσειρά του πλήθους εκδόσεων ανά έτος μεταξύ 1936 και 2019.

Μεθοδολογία

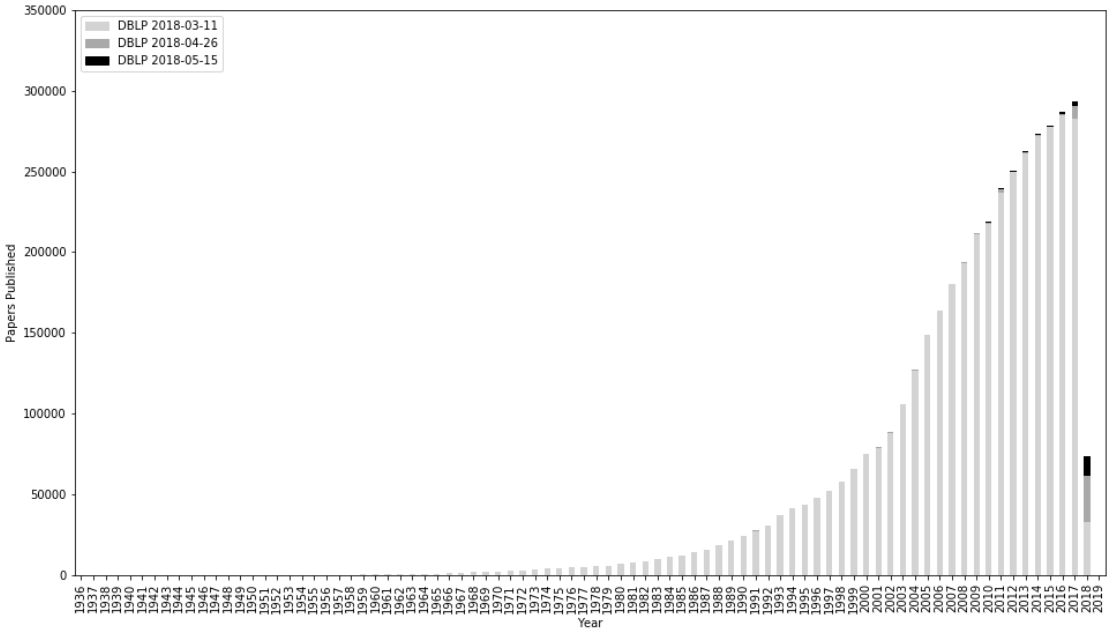
Στη δεδομένη χρονοσειρά δημοσιεύσεων θα γίνει ανάλυση με στόχο την εξαγωγή χρήσιμων χαρακτηριστικών τα οποία θα βοηθήσουν στην πρόβλεψη τιμών για τα τελευταία/προσεχή έτη. Η συγκεκριμένη διεργασία αποτελείται από τρία σκέλη: ανάγνωση αρχείου, ανάλυση δεδομένων και πρόβλεψη τιμών. Στο πρώτο σκέλος θα γίνει πρώτη επαφή και προσπάθεια ανάγνωσης ενός άγνωστης δομής αρχείου, στο δεύτερο θα γίνει οπτικοποίηση μεγεθών και αναγνώριση χαρακτηριστικών της χρονοσειράς, και στο τρίτο θα γίνει αξιολόγηση και επιλογή βέλτιστων μεθόδων πρόβλεψης για μελλοντικές τιμές.

Η εκτέλεση της ανάλυσης επιλέχθηκε να γίνει με Python και R, λόγω των εκτενών βιβλιοθηκών που περιλαμβάνουν για στατιστική και εξόρυξη δεδομένων. Το μεγαλύτερο μέρος της εργασίας προγραμματίστηκε ταυτόχρονα στις δυο γλώσσες: η Python επιλέχθηκε αρχικά σε πιο τεχνικές διεργασίες όπως ανάγνωση, αποθήκευση και οπτικοποίηση των δεδομένων, ενώ η R χρησιμοποιήθηκε για την επεξεργασία, εξόρυξη και πρόβλεψη. Η τελική υλοποίηση έγινε εξολοκλήρου σε Python και χρησιμοποιήθηκαν αντίστοιχες βιβλιοθήκες με αυτές της R για τη μεταφορά των αλγορίθμων.

Ανάγνωση αρχείου

Λόγω του μεγέθους του αρχείου και περιορισμών στη μνήμη του συστήματος, δεν είναι εφικτή η ανάγνωσή του με συνήθη προγράμματα όπως Chrome, Notepad, κλπ, αλλά θα γίνει χρήση ενός Python script που δεν απαιτεί φόρτωμα όλου του αρχείου στη μνήμη και θα επιστρέψει από το αρχείο μόνο το ζητούμενο header. Από τις πρώτες 20 σειρές μπορεί να γίνει εξοικείωση με την εσωτερική δομή του αρχείου, όπου αναγνωρίζονται τα XML tags κάθε άρθρου (ArticleID) που περικλείουν tags πληροφοριών όπως συγγραφέας, τίτλος, έτος, URL, μεταξύ άλλων.

Διάγραμμα 1: Πλήθος δημοσιεύσεων ανά έτος για τρεις εκδόσεις αρχείου εισόδου DBLP



Στη συνέχεια, με τη βοήθεια μιας μεθόδου εξαγωγής αναζητώνται όλες οι σειρές που περιέχουν το tag <year> και αποσπάται το έτος το οποίο έπειτα προστίθεται σε λίστα με όλα τα έτη δημοσιεύσεων. Η λίστα αθροίζεται ανά έτος, μετατρέπεται σε χρονοσειρά συνόλων δημοσιεύσεων για τα έτη 1936-2019, και αποθηκεύεται σε αρχείο μορφής CSV ώστε οι μετέπειτα διεργασίες να μην απαιτούν πολλαπλούς ίδιους επανυπολογισμούς και να εξοικονομείται υπολογιστικός χρόνος.

Η διαδικασία αυτή ακολουθήθηκε για την αρχική δοσμένη ημερομηνία (11/03/18), αλλά και για δυο μετέπειτα ημερομηνίες (26/04/2018 & 15/05/2018) ώστε να διαπιστωθεί ποιοτικά ο ρυθμός ανανέωσης των αρχικών δεδομένων. Οι τρεις χρονοσειρές απεικονίζονται γραφικά στο Διάγραμμα 1 όπου παρατηρούνται τα εξής χαρακτηριστικά:

- Κατά το Β'ΠΠ γίνονταν λιγότερες από 10 δημοσιεύσεις ετησίως, ενώ μόνο στο 2016 έγιναν σχεδόν 285.000.
- Συνεχής ομαλή αυξητική τάση: εκθετική (κοίλα άνω) μέχρι το 2002, λογαριθμική (κοίλα κάτω) από 2002 και έπειτα.
- Σημαντική διαφορά τιμών μεταξύ των αρχείων διαφορετικών ημερομηνιών για τα έτη 2017 (3%), 2018 (85%), και 2019 που στο παλαιότερο αρχείο εσφαλμένα έχει μια δημοσίευση. Έτη πριν το 2016 έχουν απόκλιση μικρότερη του 0.3%.

Προσέγγιση καμπύλης

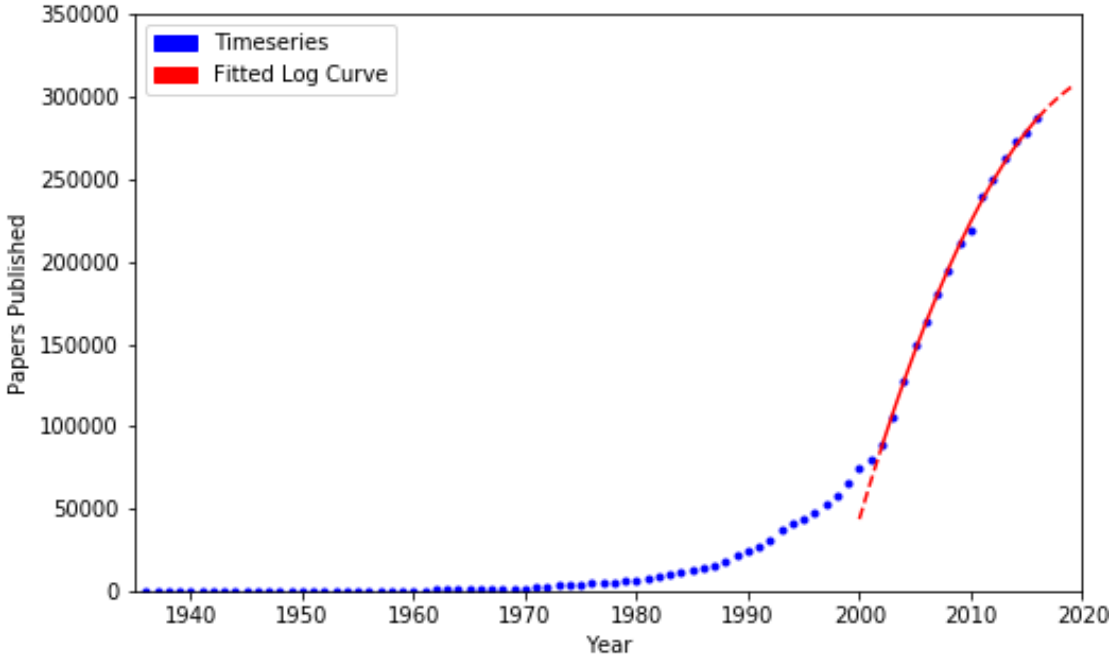
Από το Διάγραμμα 1 είναι διακριτή η λογαριθμική αυξητική τάση από το 2002 και έπειτα, οπότε δοκιμάστηκε η προσέγγιση με λογαριθμική καμπύλη για να δοθεί μια προσεγγιστική μαθηματική μορφή στο διάγραμμα που ενδεχομένως βοηθήσει στην περαιτέρω ανάλυση και πρόβλεψη. Εφαρμόζοντας τη μέθοδο Curve Fitting προκύπτει μια καμπύλη που, οπτικά τουλάχιστον, προσεγγίζει εξαιρετικά τη χρονοσειρά, όπως φαίνεται στο Διάγραμμα 2. Επίσης, θα μπορούσε να γίνει δοκιμή προσαρμογής εκθετικής καμπύλης στη χρονοσειρά για τα έτη πριν το 2002, με παρόμοιο αναμενόμενο αποτέλεσμα. Η λογαριθμική καμπύλη που προέκυψε είναι της μορφής $y = a\ log(x) + b\ x + c$, με παραμέτρους $(a, b, c) = (3.8 \cdot 10^9, -1.9 \cdot 10^6, -2.5 \cdot 10^{10})$ και $RMSE = 1774.58$.

Πρόβλεψη

Σύγκριση μεθόδων πρόβλεψης

Από τις παρατηρήσεις του Διαγράμματος 1 μπορούν να εξαχθούν δυο συμπεράσματα. Πρώτον, κατά τα έτη γύρω στο 2002 έλαβε χώρα κάποιο ιστορικής σημασίας γεγονός (όχι απαραίτητα στο χώρο του CS) που προκάλεσε μεταβολή στην αυξητική τάση των δημοσιεύσεων, το οποίο μπορεί να διασταυρωθεί ποιοτικά με το χρονοδιάγραμμα του CS της Wikipedia (Wikipedia.org, 2017). Δεύτερον, δεν έχει ολοκληρωθεί η διαδικασία δημοσιεύσεων για το 2017 και το 2018 καθώς είναι το τρέχον έτος. Συνεπώς, η ανάλυση θα στηριχθεί στα έτη 2002-2016 που εμφανίζουν την απαιτούμενη στατικότητα και θα χρησιμοποιηθούν ως σύνολα εκπαίδευσης και δοκιμής για τους αλγόριθμους πρόβλεψης που θα εξάγουν τις τιμές των ετών 2017-2019. Προβλέψεις για μετέπειτα έτη μπορούν να βασιστούν στην ίδια μέθοδο που ακολουθείται για το 2019.

Διάγραμμα 2: Προσέγγιση τμήματος χρονοσειράς με λογαριθμική καμπύλη



Για να επιλεχθεί κατάλληλη μέθοδος πρόβλεψης για τα έτη 2017-2019 πρέπει πρώτα να γίνει σύγκριση των υποψηφίων μεθόδων βάσει ενός συνόλου δοκιμής, το οποίο θα είναι τα τελευταία έτη των ετών 2002-2016. Καθώς η αναλογία συνόλου εκπαίδευσης/δοκιμής είναι συνήθως 4:1, επιλέγονται τα έτη 2002-2013 ως σύνολο εκπαίδευσης των μεθόδων πρόβλεψης, και τα έτη 2014-2016 ως σύνολο δοκιμής. Παρατηρείται ότι η χρονοσειρά χαρακτηρίζεται από τάση, αλλά δεν εμφανίζει περιοδικότητα, οπότε επιλέγονται αντίστοιχες μέθοδοι. Οι μέθοδοι πρόβλεψης που θα συγκριθούν είναι οι εξής:

1. Προσέγγιση με λογαριθμική συνάρτηση (Log Curve Fit): πρόβλεψη εντός διαστήματος εμπιστοσύνης βάσει παρατηρήσεων
2. Απλοϊκή (Naive): πρόβλεψη ίση με τελευταία παρατήρηση
3. Μέσου όρου (Average): πρόβλεψη ίση με μέσο όρο όλων των παρατηρήσεων
4. Κινούμενου μέσου όρου (Moving average): πρόβλεψη ίση με μέσο όρο παρατηρήσεων τελευταίας περιόδου (Singh, 2018)
5. Απλής εκθετικής εξομάλυνσης (Simple Exponential Smoothing): πρόβλεψη ίση με μέσο όρο παρατηρήσεων βάσει εκθετικού συντελεστή βαρύτητας
6. Γραμμικής Τάσης Holt (Holt's linear trend): πρόβλεψη βάσει τάσης από αποσύνθεση χρονοσειράς (Kalekar, 2004)
7. Αυτορρυθμιζόμενου κινούμενου μέσου όρου (ARIMA) με αυτόματη αναζήτηση συντελεστών (p,d,q): πρόβλεψη βάσει περιοδικότητας και κινούμενου μέσου όρου παρατηρήσεων (Chatfield, 2000)

Οι προβλέψεις των αλγορίθμων για τα έτη δοκιμής αναπαρίστανται παράλληλα με τις πραγματικές τιμές των ετών 2014-2016 στο Διάγραμμα 3. Είναι φανερό ότι οι πλησιέστερες ευθείες στο σύνολο δοκιμής (διακεκομμένη) λαμβάνονται από την προσέγγιση με λογαριθμική συνάρτηση και τη μέθοδο ARIMA με παραμέτρους (0,1,0) το οποίο επιβεβαιώνεται από τα σφάλματα (RMSE) των μεθόδων που παρουσιάζονται σε μειούμενη σειρά στον Πίνακα 1.

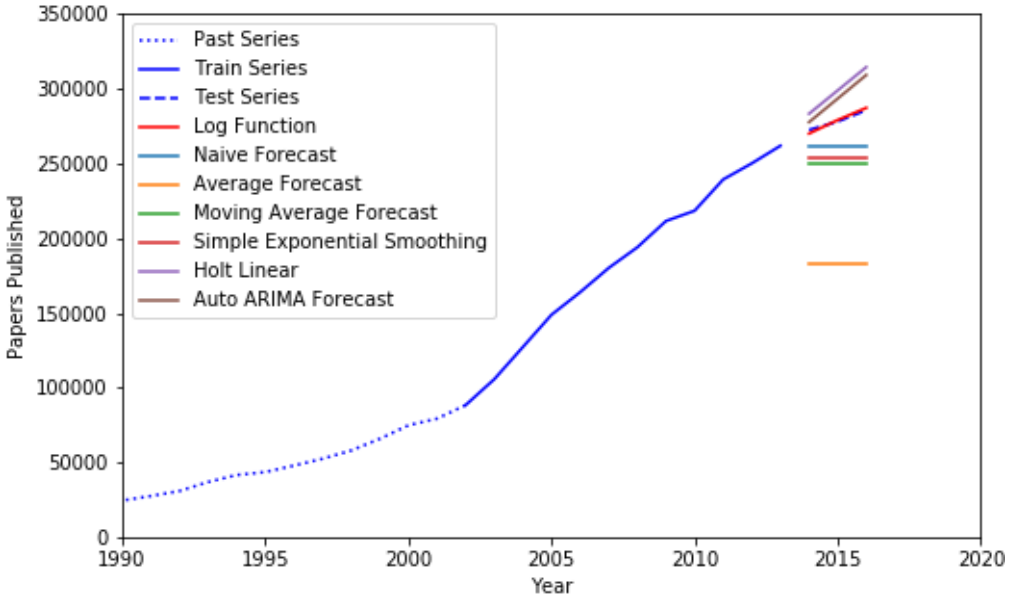
Πίνακας 1: Σύγκριση σφαλμάτων πρόβλεψης

Μέθοδος Πρόβλεψης	RMSE
Average Forecast	96370.48
Moving Average Forecast	28772.65
Exponential Smoothing	25630.37
Holt Linear	21458.62
Naive Forecast	17671.90
Auto ARIMA Forecast	16504.07
Log Curve Fit	1774.58

Τελική πρόβλεψη

Βάσει των δυο μεθόδων που ξεχώρισαν από την προηγούμενη σύγκριση, καθώς εμφάνισαν τα χαμηλότερα σφάλματα για τη δεδομένη χρονοσειρά, μπορεί να γίνει πρόβλεψη και αποτίμηση αυτής για τα έτη ενδιαφέροντος 2017-2019. Ως σύνολο εκπαίδευσης τίθεται ολόκληρο το σύνολο 2002-2016 και οι προβλέψεις με τα διαστήματα εμπιστοσύνης (95%/4σ) τους παρουσιάζονται στο Διάγραμμα 3. Παρατηρείται ότι το διάστημα εμπιστοσύνης του ARIMA αυξάνεται με την εξέλιξη των ετών, το οποίο οφείλεται στο ότι οι μελλοντικές τιμές εξαρτώνται από παλαιότερες, ενώ στην περίπτωση της λογαριθμικής προσέγγισης το διάστημα εμπιστοσύνης είναι συνάρτηση της διασποράς του συνόλου εκπαίδευσης, το οποίο είναι σταθερή τιμή. Οι τιμές των προβλέψεων για τα έτη 2017-2019 παρουσιάζονται στον Πίνακα 2.

Διάγραμμα 3: Σύγκριση μεθόδων πρόβλεψης με σύνολο δοκιμής 2014-2016



Πίνακας 2: Τιμές προβλέψεων 2017-2019

Έτος	ARIMA	Log Fit
2017	299621	294287
	min:289385	±3549
	max:309858	
2018	313704	300530
	min:299227	±3549
	max:328180	
2019	327786	305835
	min:310056	±3549
	max:345516	

Βάσει των αποτελεσμάτων αυτών θα μπορούσε κανείς να καταλήξει ότι η βέλτιστη μέθοδος πρόβλεψης για τη συγκεκριμένη χρονοσειρά είναι η προσέγγιση με λογαριθμική συνάρτηση συγκριτικά με την ARIMA, καθώς παρουσιάζει τη μικρότερη απόκλιση/σφάλμα. Αυτό μπορεί να είναι αλήθεια στο συγκεκριμένο παράδειγμα, αλλά πρέπει να τονιστεί πως η μέθοδος προσέγγισης αδυνατεί να λάβει υπόψιν μεταβολές στη συμπεριφορά της χρονοσειράς που μπορεί να οφείλονται σε εξωτερικούς παράγοντες. Απαιτείται ανεπτυγμένη διαίσθηση για το ποια τμήματα μιας χρονοσειράς και σε ποιο βαθμό μπορούν να προσεγγιστούν με μια από τις βασικές μορφές συναρτήσεων. Συνεπώς, ένα συμπέρασμα είναι πως η βέλτιστη μέθοδος πρόβλεψης για μια χρονοσειρά εξαρτάται σημαντικά από τη συμπεριφορά της ίδιας της χρονοσειράς.

Μεθοδολογία για Μεγάλα Δεδομένα

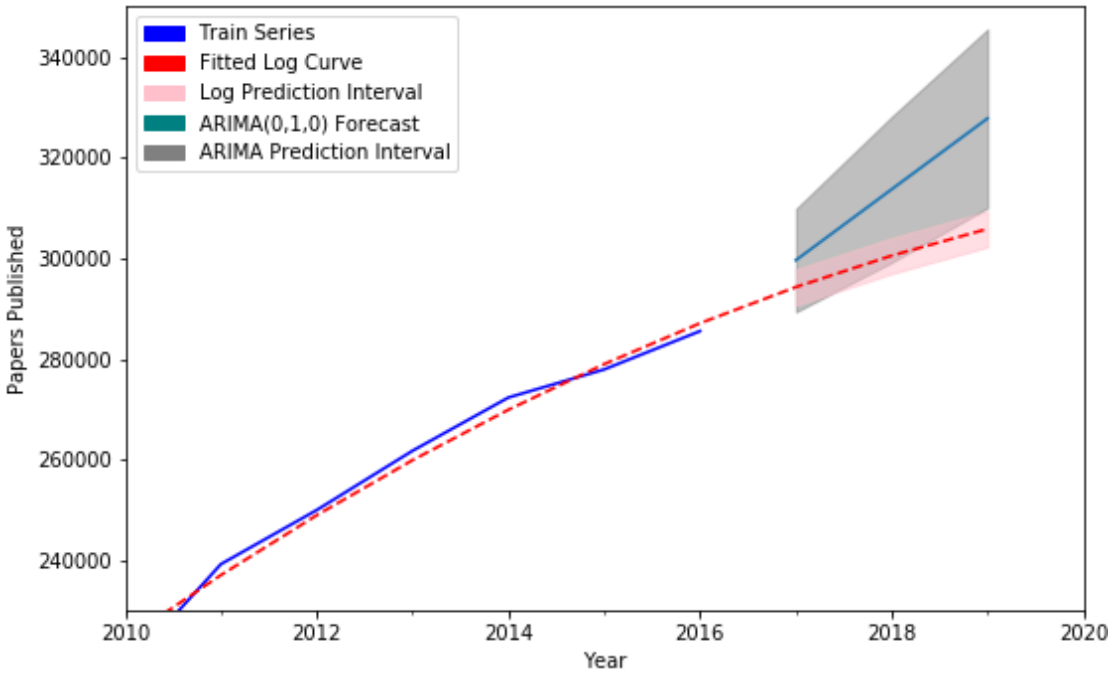
Για δεδομένα πολλαπλάσιου μεγέθους το πρόβλημα που καλείται να επιλύσει ο αναλυτής βρίσκεται κυρίως στην ανάγνωση του αρχείου εισόδου. Το στάδιο ανάλυσης και πρόβλεψης της χρονοσειράς στο συγκεκριμένο πρόβλημα γίνεται με ένα διάνυσμα μικρού μεγέθους και λίγων εγγραφών-αθροισμάτων, με αλγόριθμους χαμηλής πολυπλοκότητας και απαιτήσεων, συνεπώς δεν παρουσιάζει δυσκολία. Η διεργασία που θα απαιτήσει την επιπλέον επεξεργαστική ισχύ μετά από την αρχική αναγνώριση του header του αρχείου είναι η ανάγνωση ολόκληρου του αρχείου προς επεξεργασία.

Η σειριακή ανάγνωση ενός αρχείου μεγάλου μεγέθους έχει επίπτωση στο χρόνο που απαιτείται για την εξαγωγή του τελικού αποτελέσματος. Καθώς για ένα δεδομένο σύστημα από ορισμένο μέγεθος αρχείου και έπειτα το κόστος χρόνου θα ξεπεράσει τα αποδεκτά πλαίσια, η ανάγνωση και επεξεργασία θα πρέπει να γίνουν σε ένα παράλληλο υπολογιστικό σύστημα (Hadoop, Spark, Dask, κλπ) με τη βοήθεια αλγορίθμων παράλληλης επεξεργασίας όπως το MapReduce. Για την υλοποίηση μιας τέτοιας διαδικασίας αρχικά θα προτιμηθεί το Spark λόγω ταχύτητας, αλλά εάν το μέγεθος των δεδομένων ξεπερνάει το μέγεθος μνήμης του συστήματος η υλοποίηση θα πρέπει να γίνει σε Hadoop που μπορεί να τα αποθηκεύσει και να τα επεξεργαστεί στο HDFS (Zaharia, 2010). Η υλοποίηση του αλγορίθμου MapReduce θα είναι παρόμοια με το γνωστό παράδειγμα WordCount, με tuples της μορφής <year, paper> (Dean, 2008).

Σύνοψη

Στην εργασία αυτή έγινε αναγνώριση και ανάγνωση ενός άγνωστου αρχείου δεδομένων. Τα δεδομένα αυτά μετατράπηκαν σε χρονοσειρά, στις τιμές της οποίας έγινε ανάλυση και πρόβλεψη.

Διάγραμμα 4: Καμπύλες πρόβλεψης και διαστήματα εμπιστοσύνης 2017-2019



Έγινε σύγκριση διαφόρων μεθόδων πρόβλεψης και επιλογή των δυο βέλτιστων, βάσει των οποίων έγινε τελική πρόβλεψη και αποτίμηση των διαστημάτων εμπιστοσύνης. Τέλος, μελετήθηκε η επίπτωση του μεγέθους των δεδομένων εισόδου σχετικά με τους περιορισμούς ενός συστήματος πάνω στον υπολογιστικό χρόνο και προτάθηκαν λύσεις.

Εν κατακλείδι, μπορούν να εξαχθούν τα εξής συμπεράσματα για έναν αναλυτή που επιχειρεί να διερευνήσει/αναλύσει ένα σύνολο δεδομένων:

- Διαφορετικές υπολογιστικές μέθοδοι δεν εξάγουν απαραίτητα τα ίδια αποτελέσματα. Απαιτείται δοκιμή και επαλήθευση με πολλαπλές μεθόδους για την εξασφάλιση έγκυρων αποτελεσμάτων.
- Δεν υπάρχει μια γλώσσα προγραμματισμού ή ένα πακέτο λογισμικού που να παρέχει όλες τις λειτουργίες ή την ευκολία χρήσης που χρειάζεται ο αναλυτής. Πολύπλευρη εξοικείωση με διάφορες γλώσσες και πακέτα εξασφαλίζει υψηλότερη απόδοση.
- Το μέγεθος των δεδομένων υπαγορεύει την ισχύ και δομή του υπολογιστικού συστήματος. Μεγάλα δεδομένα απαιτούν κατανεμημένα παράλληλα συστήματα, αλλά λίγα δεδομένα μπορεί να έχουν καλύτερη απόδοση σε τοπικό σειριακό.

Βιβλιογραφία

Chatfield, C. (2000). *Time-series forecasting*. CRC Press.

Dean, J. a. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, pp. 107-113.

Hartwig, F. a. (1979). *Exploratory data analysis. Vol. 16*. Sage.

Kalekar, P. S. (2004). Time series forecasting using holt-winters exponential smoothing. *Kanwal Rekhi School of Information Technology*, pp. 1-13.

NIST/SEMATECH. (n.d.). *Engineering Statistics Handbook*. Retrieved from What is EDA?: https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm

Singh, G. (2018, 02 08). *Analytics Vidhya*. Retrieved from https://www.analyticsvidhya.com/blog/2018/02/time-series-forecasting-methods/

Tukey, J. W. (1977). *Exploratory data analysis. Vol. 2.*

University of Trier. (n.d.). *Computer Science Bibliography*. Retrieved from https://dblp.uni-trier.de/

Wikipedia.org. (2017, 10). *Timeline of computing*. Retrieved from https://en.wikipedia.org/wiki/Timeline_of_computing#Graphical_timeline

Zaharia, M. e. (2010, 10 10). Spark: Cluster computing with working sets. *HotCloud* , p. 95.