Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Μοντέλα και πρόβλεψη συμπεριφοράς χρηστών σε δημοπρασίες επιχορηγούμενης αναζήτησης

Διπλωματικη Εργασια

ΣΤΑΜΟΣ ΦΙΛΙΠΠΟΣ

Επιβλέπων: Φωτάκης Δημήτρης
Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Αύγουστος 2018

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Μοντέλα και πρόβλεψη συμπεριφοράς χρηστών σε δημοπρασίες επιχορηγούμενης αναζήτησης

Διπλωματικη Εργασια

του

Στάμου Φίλιππου

Επιβλέπων: Φωτάκης Δημήτρης
Επίκουρος Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 27η Αυγούστου 2018.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Φωτάκης Δημήτρης	Γκούμας Γεώργιος	Παπασπύρου Νικόλαος
Επίκ. Καθηγητής	Επίκ. Καθηγητής	Αν. Καθηγητής
Ε.Μ.Π	Ε.Μ.Π	Ε.Μ.Π

Αθήνα, Αύγουστος 2018

(Υπογραφή)

.....

Σταμος Φιλιππος

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Ευχαριστίες

Οι ευχαριστίες μου προς τον κύριο Φωτάκη δεν θα είναι μόνο για την βοήθεια που μου έδωσε κατά την διάρκεια της διπλωματικής αυτής όσο για την γενικότερη στήριξη που μου προσέφερε. Με βοήθησε στις αποφάσεις σχετικά με το μέλλον μου, στήριξε τις επιλογές μου και έδειξε κατανόηση στις ανάγκες που είχα. Γι΄ αυτό του οφείλω ένα μεγάλο και ειλικρινές ευχαριστώ.

Τέλος οι ευχαριστίες σε φίλους και οικογένεια είναι αυτονόητες, αλλά ένα ειδικό ευχαριστώ στον φίλο μου τον Νίκο, που προσέφερε καθοριστική βοήθεια στο τεχνικό κομμάτι μαθαίνοντας μου τον χειρισμό πολύ χρήσιμων εργαλείων που χρησιμοποιήθηκαν στην διπλωματική αυτή.

Περίληψη

Η μελέτη της διπλωματικής αυτής εστιάζει στην πειραματική αξιολόγηση μοντέλων πρόβλεψης της συμπεριφοράς του χρήστη, κατά τις δημοπρασίες επιδοτούμενων διαφημίσεων στο διαδίκτυο.

Η πλειοψηφία των εσόδων των μεγαλύτερων διαδικτυακών εταιρειών όπως η Google ή η Yahoo προέρχεται από διαφορετικών ειδών δημοπρασίες σχετικές με διαφημίσεις προϊόντων. Ενδεικτικά, το 2005, το 98% των εσόδων της Google προήλθε από δημοπρασίες και πιο ειδικά, από δημοπρασίες GSP (generalized second prize). Στο πρώτο κομμάτι της διπλωματικής θα γίνει μελέτη των δημοπρασιών αυτών, των ιδιοτήτων τους, καθώς και θα γίνει σύγκριση τους με την κατηγορία δημοπρασιών VCG (Vickrey-Clarke-Groves).

Η διπλωματική αυτή εστιάζει στις επιχορηγούμενες διαφημίσεις κατά τις αναζητήσεις των χρηστών στο διαδίκτυο. Με στόχο την μεγιστοποίηση των κερδών και την εύρεση της βέλτιστης τοποθέτησης των διαφημίσεων, ιδιαίτερη έμφαση δίνεται στην πρόβλεψη της συμπεριφοράς του χρήστη κατά την πλοήγηση του στις μηχανές αναζήτησης. Μελετώνται τα δύο βασικά μοντέλα της βιβλιογραφίας, το Αλληλουχίας και το Διαχωρίσιμο μοντέλο.

Η ανάλυση της συμπεριφοράς του χρήστη μπορεί να πραγματοποιηθεί με χρήση Μηχανικής Μάθησης γι΄ αυτό στην συνέχεια παρουσιάζονται οι βασικές έννοιες που χρειάζονται για την ανάλυση των δεδομένων. Παρ΄ ότι εν τέλει δεν χρησιμοποιήθηκε μηχανική μάθηση λόγω των χαρακτηριστικών των δεδομένων, οι γνώσεις σχετικές με τις διαφορετικές μετρικές για την αξιολόγηση μοντέλων πρόβλεψης της συμπεριφοράς του χρήστη ήταν απαραίτητες και τεχνικές αξιολόγησης των μοντέλων χρησιμοποιήθηκαν εκτεταμένα.

Στο τελευταίο κομμάτι της διπλωματικής γίνεται πειραματική αξιολόγηση των μοντέλων αυτών. Χρησιμοποιούνται τα δεδομένα με τίτλο Personalized Web Search που δημοσιεύτηκαν από την Yandex στα πλαίσια ενός διαγωνισμού. Τα δεδομένα αυτά είναι οργανικά και γίνεται η υπόθεση ότι ο χρήστης συμπεριφέρεται με τον ίδιο τρόπο στα οργανικά και στα επιχορηγούμενα αποτελέσματα. Τα χαρακτηριστικά των δεδομένων αυτών αναλύονται εκτεταμένα με χρήση Hadoop και γίνεται αναλυτική σύγκριση των δύο βασικών μοντέλων της βιβλιογραφίας. Στο τέλος, γίνεται μία προσπάθεια εύρεσης και αξιολόγησης νέων μοντέλων. Το μοντέλο στο οποίο καταλήγουμε βρίσκεται πολύ κοντά στα ποσοστά επιτυχίας των μοντέλων της βιβλιογραφίας, με αυτό να πετυχαίνει μεγαλύτερη ακρίβεια στην πρόβλεψη των κλικ του χρήστη.

Λέξεις Κλειδιά

Σχεδιασμός Μηχανισμών, Δημοπρασίες Επιχορηγούμενων Διαφημίσεων, Μοντέλο Αλληλουχίας, Διαχωρίσιμο Μοντέλο, Μηχανική Μάθηση

Abstract

This thesis focuses on the experimental evaluation of the user's behaviour in sponsored search auctions.

The majority of income of the biggest web companies like Google and Yahoo, is earned through auctions related to the advertisements of products. For example, in 2005, 98% of Google's revenue derived from GSP (generalized second prize) auctions. In the first part of this thesis, we will study these auctions, their properties and we will compare GSP auctions with VCG (Vickrey-Clarke-Groves) auctions, another widely used type of auction.

This thesis focuses on the sponsored search auctions problem happening on the web. Aiming to maximize the company's revenue and to find the best possible allocation of ads, it becomes really important to be able to predict user's behavior while browsing with the use of search engines. The two most widely used models, the Cascade Model and the Separable Model are studied and widely analyzed.

The analysis of user's behavior can be predicted through Machine Learning. That is why we continue with knowledge on the basics of Machine Learning. Although in the end Machine Learning methods weren't used because of limited data, the understanding of metrics in order to evaluate the analyzed models was crucial during the analysis of the bibliographic models.

In the last part of this thesis, models are evaluated experimentally. Yandex's Personalized Web Search dataset is used, analyzed with the use of Hadoop. As this data is the result of organic searches, we make the hypothesis that users behave the same way when being given organic and sponsored results. Firstly, the two main bibliography models are analyzed and compared extensively and in the end new models are tried. The model that we propose has similar accuracy with the other known models but is able to perform better at predicting when users click, while the rest of the models perform better at predicting when users don't click.

Keywords

Sponsored Search Auctions, Mechanism Design, Generalized Second-Price Auctions, Vickrey-Clarke-Groves Auctions, Cascade Model, Separable Model, Machine Learning, Hadoop

Περιεχόμενα

Κατάλογος Σχημάτων

Κεφάλαιο 1

Εισαγωγή

Η θεωρία παιγνίων και ο σχεδιασμός μηχανισμών, είναι μία επιστήμη που αφορά διαφορετικούς τομείς της ζωής και μελετάται από μηχανικούς υπολογιστών, οικονομολόγους, κοινωνιολόγους κ.α. Ένας απ΄ τους βασικούς λόγους ραγδαίας επιστημονικής εξέλιξης και συστηματικής μελέτης των τομέων αυτών είναι ότι οι οικονομολόγοι επιχειρούν να προβλέψουν πως η αγορά θα διαμορφωθεί. Με στόχο την μεγιστοποίηση του κέρδους, επιχειρούν να δημιουργήσουν μηχανισμούς πληρωμών των οποίων οι τιμές ισορροπίας να διαμορφώνονται όσο το δυνατόν ψηλότερα.

Στα πλαίσια των παραπάνω, ξεχωριστό κλάδο αφορούν οι δημοπρασίες. Στην σύγχρονη κοινωνία όπου ο ανταγωνισμός βρίσκεται στα ύψη, η πώληση αγαθών είτε υπηρεσιών προσφέρονται με δημοπρασίες. Στις δημοπρασίες αυτές, το άτομο που θα βγει νικητής πληρώνει ένα αντίστοιχο αντίτιμο και με αυτόν τον τρόπο κερδίζει το προϊόν. Το αντίτιμο αυτό αποτελεί έναν πολύ σημαντικό κλάδο του σχεδιασμού μηχανισμών, ο οποίος ασχολείται με την μεγιστοποίηση των κερδών του δημοπράτη. Με την ανακάλυψη δύο πολύ σημαντικών και πλούσιων σε ιδιότητες μηχανισμών, τους GSP (Γενικευμένες δημοπρασίες δεύτερης τιμής) και τις δημοπρασίες VCG (Vickrey-Clarke-Groves) ο κλάδος αυτός έχει ανθίσει.

Ενδεικτικά, τα έσοδα σχεδόν εξ΄ ολοκλήρου των γιγαντιαίων εταιρειών του διαδικτύου όπως η Google προέρχονται από δημοπρασίες. Όταν εμπλέκονται δισεκατομμύρια σε έναν κλάδο της επιστήμης, όπως είναι αναμενόμενο γίνονται εκτεταμένες έρευνες για την βελτιστοποίηση των μηχανισμών αυτών. Οι δημοπρασίες επιχορηγούμενων αναζητήσεων, είναι μία κατηγορία δημοπρασιών που έχει ιδιαίτερη σημασία για τις εταιρείες αυτές. Οι δημοπρασίες αυτές πραγματοποιούνται με σκοπό την πώληση μίας διαφημιστικής θέσης στις σελίδες αποτελεσμάτων των χρηστών που αναζητούν την λέξη που τους ενδιαφέρει. Επομένως, όταν ο χρήστης πληκτρολογήσει την αναζήτησή του σε μία μηχανή αναζήτησης, οι νικητές των δημοπρασιών σχετικά με τις φράσεις-λέξεις θα βρίσκονται στις πρώτες θέσεις των αποτελεσμάτων και δίπλα σε αυτούς θα αναγράφεται ότι είναι χρηματοδοτούμενες διαφημίσεις.

Στην κατηγορία των δημοπρασιών αυτών, εξαιρετικά σημαντικό ρόλο έχει η επιτυχής πρόβλεψη της συμπεριφοράς του χρήστη, δηλαδή την επιλογή του σε ποιά URLs θα κάνει click. Η σωστότερη πρόβλεψη της συμπεριφοράς του χρήστη μπορεί να οδηγήσει σε βέλτιστες τοποθετήσεις των διαφημίσεων στα πλαίσια της μεγιστοποίησης των κερδών. Δύο βασικά

μοντέλα είναι γνωστότερα στην βιβλιογραφία, το Αλληλουχίας και το Διαχωρίσιμο τα οποία επιχειρούν με απλό τρόπο να προβλέψουν τις κινήσεις του χρήστη.

1.1 Αντικείμενο της διπλωματικής

Η διπλωματική αυτή εστιάζει στην πειραματική αξιολόγηση των 2 μοντέλων που προαναφέρθηκαν για την πρόβλεψη της συμπεριφοράς του χρήστη. Για την μελέτη αυτή, προαπαιτούνται γνώσεις από διαφορετικούς επιστημονικούς κλάδους.

Η πλήρης και σε βάθος κατανόηση των δημοπρασιών, προϋποθέτει την μελέτη βασικών εννοιών της θεωρίας παιγνίων και του σχεδιασμού μηχανισμών, καθώς βοηθά στην κατανόηση των δημοπρασιών, των κινήτρων των παιχτών που εμπλέκονται σε αυτές, καθώς και στην κατανόηση του τρόπου με τον οποίο διαμορφώνονται οι ισορροπίες στην αγορά.

Ξεχωριστά μελετώνται τα είδη δημοπρασιών και ειδικότερα οι δημοπρασίες GSP [10, 16, 6] , οι οποίες χρησιμοποιούνται σχεδόν αποκλειστικά στις δημοπρασίες επιχορηγούμενων διαφημίσεων. Τα μοντέλα πρόβλεψης της συμπεριφοράς του χρήστη [17, 14, 18, 27, 19] και οι δημοπρασίες GSP συνδέονται άρρηκτα, καθώς έχουν τον ίδιο σκοπό, την μεγιστοποίηση των κερδών, αλλά καθώς και γιατί ο ένας μηχανισμός υποβάλλει περιορισμούς στον άλλον, καθώς διαφορετικά μοντέλα πρόβλεψης της συμπεριφοράς του χρήστη ενδεχομένως να οδηγούν σε διαφορετικές τιμές κατά τις δημοπρασίες.

Για την ανάλυση δεδομένων, εξαιρετικά χρήσιμες είναι οι γνώσεις μηχανικής μάθησης [33, 1], γιάυτό εισάγονται βασικές μέθοδοι αυτής. Μάλιστα, παρουσιάζονται εκτεταμένα διαφορετικές μετρικές και τρόποι αξιολόγησης μοντέλων πρόβλεψης, τα οποία χρησιμοποιούνται στην αξιολόγηση των μοντέλων της βιβλιογραφίας.

Η ανάλυση των δεδομένων γίνεται χρησιμοποιώντας Hadoop [8, 31, 7], προγραμματίζοντας στην λογική του MapReduce. Τα δεδομένα που χρησιμοποιούνται με τίτλο Personalized Web Search που δημοσιεύτηκαν από την Yandex στα πλαίσια ενός διαγωνισμού, είναι οργανικά αποτελέσματα και όχι επιχορηγούμενων δημοπρασιών. Λόγω της δυσκολίας εύρεσης των δεδομένων που ήταν απαραίτητα, γίνεται η υπόθεση ότι ο χρήστης συμπεριφέρεται στα οργανικά αποτελέσματα με τον ίδιο τρόπο με τον οποίο συμπεριφέρεται στα επιχορηγούμενα. Τα δεδομένα οργανώνονται, αναλύονται και σε αυτά εφαρμόζονται το μοντέλο Αλληλουχίας και το Διαχωρίσιμο μοντέλο. Γίνεται σύγκριση και αξιολόγηση αυτών αλλά δοκιμάζονται και καινούρια μοντέλα. Από αυτά, προτείνουμε ένα καινούριο μοντέλο που ανταγωνίζεται τα μοντέλα της βιβλιογραφίας και αποδίδει ικανοποιητικότερα σε ορισμένες περιπτώσεις.

1.2 Οργάνωση του τόμου

Η διπλωματική αυτή είναι οργανωμένη σε 6 κεφάλαια.

Στο **Κεφάλαιο 1** γίνεται εισαγωγή στο πρόβλημα και παρουσιάζεται η συνολική εικόνα της διπλωματικής.

Στο **Κεφάλαιο 2** γίνεται εισαγωγή στην θεωρία παιγνίων και στον σχεδιασμό μηχανισμών. Στην συνέχεια παρουσιάζονται οι βασικές δημοπρασίες, ενώ εστιάζουμε στις δημοπρα-

σίες GSP και αναφέρονται ιδιότητες τους.

Στο **Κεφάλαιο 3** παρουσιάζεται αναλυτικά το πρόβλημά μας, η εφαρμογή των δημοπρασιών GSP στο πρόβλημά μας και γίνεται σύγκριση τους με τις δημοπρασίες VCG. Στην συνέχεια παρουσιάζονται τα βασικά μοντέλα πρόβλεψης της συμπεριφοράς του χρήστη, το μοντέλο Αλληλουχίας και το Διαχωρίσιμο.

Στο **Κεφάλαιο 4** γίνεται εισαγωγή στην Μηχανική Μάθηση, παρουσιάζονται λόγοι για τους οποίους η Μηχανική Μάθηση μπορεί να συμβάλει στο πρόβλημά μας και παρουσιάζονται μετρικές και τρόποι αξιολόγησης μοντέλων που προβλέπουν την συμπεριφορά του χρήστη. Στο τέλος γίνεται μία μικρή εισαγωγή σε ένα εργαλείο ανάλυσης μεγάλου όγκου δεδομένων με χρήση Μηχανικής Μάθησης.

Στο **Κεφάλαιο 5** παρουσιάζονται τα δεδομένα που βρέθηκαν, γίνεται μία εισαγωγή στο εργαλείο Hadoop που χρησιμοποιήθηκε και πραγματοποιείται η προεπεξεργασία των δεδομένων. Στην συνέχεια αναλύονται τα δεδομένα, εξάγονται στατιστικά για αυτά και τελικά γίνεται αξιολόγηση διαφορετικών μοντέλων τόσο της βιβλιογραφίας όσο και της διπλωματικής αυτής, για την πρόβλεψη της συμπεριφοράς του χρήστη.

Στο **Κεφάλαιο 6** παρουσιάζεται μία σύνοψη της διπλωματικής, συμπεράσματα, καθώς και προτείνονται νέες ερευνητικές κατευθύνσεις.

Κεφάλαιο 2

Θεωρία Παιγνίων και Σχεδιασμός Μηχανισμών

Στο κεφάλαιο αυτό θα γίνει μία γενική εισαγωγή στην επιστήμη της θεωρίας παιγνίων και στον σχεδιασμό μηχανισμών. Η σε βάθος κατανόηση την δημοπρασιών που μας ενδιαφέρουν, απαιτούν γνώσεις σχετικά με τις ιδιότητες των μηχανισμών και τα κίνητρα των παιχτών. Επιπλέον θα παρουσιαστούν οι δύο βασικές κατηγορίες δημοπρασιών που μας ενδιαφέρουν (VCG & GSP), ενώ θα παρουσιαστούν τα πλεονεκτήματά τους σε βάρος εναλλακτικών μηχανισμών.

2.1 Θεωρία Παιγνίων

Η Θεωρία παιγνίων αφορά την μελέτη υποθετικών σεναρίων κατά τις οποίες 2 ή και περισσότεροι έξυπνοι και αντικειμενικοί παίχτες, καλούνται να λάβουν την καλύτερη δυνατή απόφαση, με γνώμονα το εκάστοτε προσωπικό τους συμφέρον. Ο κλάδος αυτός αφορά ποικίλους τομείς όπως τα οικονομικά, την επιστήμη των υπολογιστών, τις κοινωνικοπολιτικές επιστήμες ή την ψυχολογία.

Το διασημότερο παράδειγμα αποτελεί το δίλημμα του φυλακισμένου. Σε αυτό, 2 εγκληματίες συλλαμβάνονται για ένα έγκλημα που έπραξαν. Επειδή όμως δεν υπάρχουν αρκετά στοιχεία για να καταδικαστούν, ο εισαγγελέας τους απομονώνει ξεχωριστά σε ένα δωμάτιο και τους δίνει τις ακόλουθες επιλογές. Εάν ο ένας κατηγορούμενος ομολογήσει πως ο συνεργάτης του έπραξε το έγκλημα και ο άλλος όχι, τότε αυτός που δεν θα ομολογήσει θα καταδικαστεί για 15 χρόνια ενώ ο πρώτος θα αφεθεί ελεύθερος. Αντίθετα, αν ομολογήσουν και οι 2, τότε θα καταδικαστούν και οι 2 για 5 χρόνια ενώ αν δηλώσουν και οι 2 αθώοι, τότε θα καταδικαστούν και οι 2 για 1 χρόνο. Τα παραπάνω σενάρια φαίνονται στον παρακάτω πίνακα :

Φυλακισμένος Α / Φυλακισμένος Β	Ομολογία	Μη ομολογία
Ομολογία εγκλήματος	5,5	0,15
Μη ομολογία εγκλήματος	15,0	1,1

Σχήμα 2.1: Δίλημμα του φυλακισμένου

Υποθέσουμε πως οι 2 παίχτες αντιλαμβάνονται πλήρως το παιχνίδι, είναι αντικειμενικοί και έξυπνοι παίχτες, καθώς και δεν έχουν δυνατότητα επικοινωνίας, ούτε είχαν προσυμφωνήσει από πριν. Έτσι, ανεξάρτητα από την επιλογή του άλλου παίχτη, η κυρίαρχη στρατηγική είναι ο κάθε παίχτης να προδώσει. Αυτό συμβαίνει καθώς αν για παράδειγμα ο παίχτης Α ομολογήσει, τότε η ομολογία οδηγεί σε 5 χρόνια φυλάκισης έναντι των 15 της μη ομολογίας. Αντίστοιχα, εάν δεν προδώσει, πάλι έχουμε 0 χρόνια έναντι του 1ός. Επομένως σε κάθε περίπτωση οι παίχτες θα επιλέξουν να προδώσουν. Το παράδοξο του σεναρίου αυτού, είναι πως παρ΄ ότι οι παίχτες είχαν την επιλογή να φυλακιστούν μόνο για 1 χρόνο, η κυρίαρχη στρατηγική τους οδήγησε στην φυλάκιση για 5 χρόνια.

Στην γενική περίπτωση για να έχουμε ένα παίγνιο, θα πρέπει να έχουν οριστεί τα ακόλουθα :

- Ένα σύνολο N από n παίχτες.
- Για κάθε παίχτη i, υπάρχει ένα σύνολο S_i από στρατηγικές ανάλογα με τις απαντήσεις των υπόλοιπων $n-1$ παιχτών.
- Επίσης, για κάθε παίχτη i χρειάζεται να οριστεί μία συνάρτηση κόστους/ κέρδους η οποία αντιστοιχίζει κάθε τούπλα n στοιχείων (στρατηγική των άλλων $n-1$ παιχτών και την στρατηγική του i-οστού παίχτη) σε έναν φυσικό αριθμό ο οποίος εκφράζει το κόστος ή κέρδος που έχει ο κάθε παίχτης απ΄ την εκάστοτε έκβαση του παιγνίου.

2.1.1 Κατηγορίες Παιγνίων

Τα παίγνια μπορούν να κατηγοριοποιηθούν ανάλογα με ορισμένα τους χαρακτηριστικά. Ενδεικτικά θα αναφερθούν οι βασικές κατηγοριοποιήσεις.

- Συμμετρικά / Μη συμμετρικά παίγνια

  Ως συμμετρικά παίγνια ορίζονται τα παίγνια στα οποία η έκβαση του αποτελέσματος για τον παίχτη εξαρτάται μόνο από τις στρατηγικές που επέλεξαν οι αντίπαλοι και όχι από το ποιός παίχτης επέλεξε ποιά στρατηγική. Έτσι, το παίγνιο του φυλακισμένου για παράδειγμα, αποτελεί συμμετρικό παίγνιο.

- Παίγνια μηδενικού ή μη αθροίσματος

  Τα παίγνια τα οποία αθροίζουν στο μηδέν είναι αυτά στα οποία το κέρδος του ενός παίχτη προκύπτει από το κόστος ενός αντίπαλου παίχτη. Στα παίγνια αυτά, θα πρέπει

το άθροισμα κέρδους και κόστους όλων των παιχτών να είναι ακριβώς μηδέν. Δημοφιλέστερο παιχνίδι που αθροίζει στο μηδέν είναι το πόκερ. Αντίθετα, το παίγνιο του φυλακισμένου δεν αθροίζει στο μηδέν.

- Συνεργατικά / Μη συνεργατικά παίγνια

  Συνεργατικά παίγνια είναι τα παίγνια στα οποία η τελική στρατηγική προκύπτει μετά από συνεργασία και διαπραγματεύσεις ανάμεσα σε όλους τους παίχτες. Αντίθετα στα μη συνεργατικά, οι παίχτες αποφασίζουν μόνοι τους προσπαθώντας να μεγιστοποιήσουν/ ελαχιστοποιήσουν την συνάρτηση κέρδους/ κόστους. Το παίγνιο του φυλακισμένου αποτελεί μη συνεργατικό παίγνιο.

- Ακολουθιακά / Παράλληλα παίγνια

  Στα ακολουθιακά παίγνια ο κάθε παίκτης αποφασίζει την στρατηγική του στην σειρά του, όσο οι υπόλοιποι παίκτες τον περιμένουν. Αντίθετα στα παράλληλα, οι παίκτες αποφασίζουν όλοι την ίδια χρονική στιγμή. Έτσι, ενώ στα ακολουθιακά οι παίκτες έχουν γνώση της επιλογής του αντιπάλου, στα παράλληλα, οι παίχτες αποφασίζουν χωρίς αυτήν την πληροφορία.

- Παίγνια Πλήρους πληροφόρησης / Παίγνια Μερικούς πληροφόρησης

  Ο διαχωρισμός αυτός αφορά την γνώση ή μη των στρατηγικών και των συναρτήσεων κόστους/κέρδους των αντίπαλων παικτών. Έτσι, ενώ το σκάκι αποτελεί παίγνιο πλήρους πληροφόρησης, το πόκερ αποτελεί παίγνιο μερικής πληροφόρησης.

2.1.2 Ισορροπίες Nash

Οι ισορροπίες Nash αφορούν την καλύτερη δυνατή επιλογή απ΄ όλους τους παίκτες. Με μοναδικό στόχο την μεγιστοποίηση του κέρδους τους/ ελαχιστοποίηση του κόστους τους, οι παίκτες θα προσπαθήσουν να επιλέξουν την καλύτερη δυνατή στρατηγική γι΄ αυτούς. Σύμφωνα με την Θεωρία Παιγνίων, κάθε παίγνιο οδηγείται σε μία ισορροπία και επομένως στόχος της επιστήμης αυτής αποτελεί η εύρεση των ισορροπιών αυτών, καθώς και η ανάλυση των ιδιοτήτων τους.

Ανάλογα με το πρόβλημα, μπορούν να υπάρξουν διαφορετικού είδους ισορροπίες. Συγκεκριμένα, υπάρχουν οι ακόλουθες κατηγορίες [24] :

- **Αμιγής Στρατηγική Ισορροπία Nash**

  Αμιγής ισορροπία Nash είναι η κατάσταση στην οποία όλοι οι παίχτες έχουν καταλήξει σε μία στρατηγική η οποία είναι η καλύτερη δυνατή γι΄ αυτούς. Έτσι, κανένας απ΄ αυτούς δεν έχει κίνητρο να αλλάξει στρατηγική και επομένως καταλήγουμε σε μία σταθερή ισορροπία. Πιο φορμαλιστικά ο ορισμός της αμιγούς ισορροπίας Nash είναι [24] :

  Ορισμός 2.1. ***Αμιγής Στρατηγική Ισορροπίας Nash (Pure Strategy Nash Equilibrium)***

  $$u_i(s_i, s_{-i}) \geq u_i(s_i', s_{-i})$$

Επομένως δεν υπάρχει στρατηγική s_i' η οποία να είναι καλύτερη από την βέλτιστη στρατηγική s_i , την οποία τελικά ο παίχτης θα συνεχίζει να παίζει σταθερά.

- **Mixed Στρατηγική Ισορροπίας Nash (Mixed Strategy Nash Equilibrium)**

  Σε ορισμένες περιπτώσεις η καλύτερη δυνατή στρατηγική είναι ο παίχτης να εναλλάσσεται σταθερά ανάμεσα σε 2 ή και περισσότερες στρατηγικές. Για παράδειγμα στο παιδικό παιχνίδι πέτρα-ψαλίδι-χαρτί, αποδεικνύεται πως η καλύτερη δυνατή στρατηγική για έναν παίχτη είναι να παίζει με πιθανότητα 1/3 το καθένα απ΄ τα 3 αντικείμενα.

  Μάλιστα αποδεικνύεται το παρακάτω θεώρημα [24] :

  Θεώρημα 2.1. *Για κάθε παιχνίδι με πεπερασμένο αριθμό παικτών, υπάρχει μία ισορροπία Nash , μεικτών στρατηγικών.*

- **Συσχετιστική Ισορροπία Nash (Correlated Nash Equilibrium)** Σε ορισμένες περιπτώσεις η ισορροπία μπορεί να επιτευχθεί μόνο εάν εμπλακεί στο παίγνιο και ένας συντονιστής. Η επεξήγηση θα δοθεί μέσω του ακόλουθου παραδείγματος. Έστω ότι διασταυρώνονται 2 αυτοκίνητα και το καθ΄ ένα έχει 2 επιλογές, είτε να σταματήσει είτε να συνεχίσει. Ο πίνακας με τις στρατηγικές τους δίνεται παρακάτω :

1 \ 2	Cross	Stop
Cross	-100 -100	0 1
Stop	1 0	0 0

Σχήμα 2.2: Διασταυρούμενα αυτοκίνητα

Εάν οι παίχτες αποφασίσουν μόνοι τους, τότε το παίγνιο έχει 3 ισορροπίες. Αυτές είναι είτε πως ένα απ΄ τα 2 αυτοκίνητα περνά, είτε οι παίχτες περνούν με πιθανότητα 1/101 και επομένως έχουν αναμενόμενο μικρό κέρδος περίπου 0.00001. Αντίθετα, εάν εμπλακεί ο συντονιστής, τότε αυτός θα αποφασίσει πως ένα απ΄ τα 2 αμάξια θα περάσει. Έτσι, αυτό θα περάσει ενώ το δεύτερο αυτοκίνητο παρ΄ ότι θα έχει κέρδος 0, δεν το συμφέρει να περάσει γιατί θα προκαλέσει ατύχημα και επομένως κόστος -100.

Αξίζει να σημειωθεί πως οι κατηγορίες αυτές έχουν σχέση υπερσυνόλων-υποσυνόλων. Πιο συγκεκριμένα : $Pure \subset Mixed \subset Correlated$

2.2 Σχεδιασμός Μηχανισμών και φιλαλήθεια

2.2.1 Κοινωνική Επιλογή

Ένα από τα συνηθισμένα προβλήματα στον σχεδιασμό ενός αθλητικού τουρνουά είναι η δημιουργία μίας δομής η οποία υποχρεώνει τις ομάδες να παίζουν πάντα για το καλύτερο δυνατό αποτέλεσμα. Αντιθέτως, είναι συχνό φαινόμενο ομάδες να αποφασίζουν πως η ήττα σε έναν αγώνα είναι προτιμότερη από την νίκη, είτε για να έχουν πιο αδύναμους μελλοντικούς αντιπάλους, είτε για να αποκλείσουν μία ανταγωνίστρια ομάδα.

Η επιστήμη του σχεδιασμού μηχανισμών και της κοινωνικής επιλογής [5, 27] αφορά την δημιουργία μηχανισμών οποιουδήποτε τομέα της ζωής, οι οποίοι προσπαθούν να επιτύχουν ορισμένες ιδιότητες όπως η μεγιστοποίηση του κοινωνικού οφέλους, την μεγιστοποίηση των κερδών του δημιουργού του μηχανισμού ή την εξασφάλιση πως όλοι οι παίκτες θα εκφράσουν τις πραγματικές τους ανάγκες και προτιμήσεις. Το παρακάτω παράδειγμα είναι ενδεικτικό των προβλημάτων που μπορεί να εμφανιστούν κατά τον σχεδιασμό ενός μηχανισμού.

Παράδειγμα Κοινωνικής Επιλογής

Έστω ότι υπάρχουν 3 υποψήφιοι στις εκλογές, ο A,B και C . Έστω επίσης πως έχουμε 1000 ψηφοφόρους για τους οποίους εάν γνωρίζαμε τις πραγματικές τους προτιμήσεις αυτές θα ήταν οι εξής :

- 499 ψηφοφόροι : $A > B > C$
- 3 ψηφοφόροι : $B > C > A$
- 498 ψηφοφόροι : $C > B > A$

Ας σκεφτούμε πως διαφορετικοί μηχανισμοί οδηγούν σε διαφορετικά αποτελέσματα. Εάν χρησιμοποιούσαμε έναν απλό μηχανισμό όπως τον απόλυτα πλειοψηφικό, τότε ο νικητής θα ήταν ο A, αφού ο μηχανισμός αυτός δεν λαμβάνει υπ΄ όψιν την 2η και 3η επιλογή. Όμως, εάν γινόταν χρήση του μηχανισμού αυτού, οι ομάδα των 3ών ψηφοφόρων, θα είχε κίνητρο να αλλάξει την ψήφο του και να επιλέξει ως πρώτη του προτίμηση τον υποψήφιο C, αφού το αποτέλεσμα των εκλογών θα ήταν πιο ευνοϊκό γι΄ αυτόν αφού τελικά θα έβγαινε νικητής ο υποψήφιος της 2ης προτίμησης τους κι όχι της τρίτης. Επομένως στο παράδειγμα αυτό ορισμένοι από τους ψηφοφόρους είχαν κίνητρο να ψηφίσουν όχι με βάση την πραγματική τους προτίμηση αλλά επηρεασμένη από τον τρόπο λειτουργία του μηχανισμού ψηφοφορίας να αλλάξουν τις προτιμήσεις που δηλώνουν.

Έστω δύο άλλοι διαφορετικοί μηχανισμοί. Σύμφωνα με τον πρώτο, η 1η ψήφος δίνει 2 πόντους, η 2η 1 πόντο και η τελευταία 0. Σύμφωνα με αυτήν ο νικητής θα έπρεπε να είναι ο C.

Ορισμός 2.1. ***Νικητής Condorcet***

Νικητής Condorcet είναι αυτός ο οποίος βγαίνει νικητής έναντι σε όλους τους υπόλοιπους υποψηφίους εάν η σύγκριση γίνει σε ζευγάρια. [13]

Σύμφωνα με τον κανόνα αυτό, νικητής Condorcet είναι ο υποψήφιος Β.

Όπως φαίνεται, 3 διαφορετικοί μηχανισμοί οδηγούν σε 3 διαφορετικούς νικητές. Μάλιστα, δεν είναι ξεκάθαρο ποιός από τους 3 είναι πιο ¨τέλειος¨, ενώ κάθε ένας απ΄ αυτούς έχεις τις αδυναμίες του και δεν επιτυγχάνει την εξασφάλιση πως οι ψηφοφόροι θα ψηφίσουν με βάση την πραγματική τους προτίμηση. Για να οριστεί ένα πλαίσιο κοινωνικής επιλογής χρειαζόμαστε τα ακόλουθα :

- Ν παίκτες.
- Μ υποψηφίους.
- Για κάθε παίκτη i, υπάρχει μία γραμμική σειρά ανάμεσα στους υποψηφίους L_i η οποία εκφράζει τις προτιμήσεις του ανάμεσα στους Μ υποψηφίους.

2.2.2 Μηχανισμοί με χρήματα και Φιλαλήθεια

Ξεχωριστή κατηγορία μηχανισμών αφορούν οι μηχανισμοί που εμπλέκουν χρήματα. Φορμαλιστικά για να οριστεί το πρόβλημα χρειαζόμαστε τα παρακάτω :

- Ένα σύνολο Ν, από n παίκτες.
- Ένα σύνολο Μ, από m εκβάσεις του αποτελέσματος.
- Ένα σύνολο συναρτήσεων V, το οποίο για κάθε παίκτη i, συμπεριλαμβάνει μία συνάρτηση $V = \{v_i : M \to \mathbb{R}\}$ η οποία εκφράζει την εκτίμηση της αξίας κάθε έκβασης γι΄ αυτόν. Συνήθως αυτή η συνάρτηση είναι άγνωστη και αποτελεί ιδιωτική πληροφορία που μόνο ο χρήστης γνωρίζει.
- Μία συνάρτηση έκβασης του αποτελέσματος $f : V^n \to M$.
- Για κάθε παίκτη i, υπάρχει μία συνάρτηση πληρωμής $p_i : V^n \to \mathbb{R}$ η οποία εκφράζει το ποσό που θα πρέπει να πληρώσει ο χρήστης για το αντικείμενο για το οποίο ενδιαφέρεται.

Ορισμός 2.2. *[27, 2]* ***Ωφέλεια χρήστη***

Η ωφέλεια u_i του χρήστη i ορίζεται ως η διαφορά της αξίας που έχει το αντικείμενο για τον χρήστη, μείον της τιμής που καλείται να πληρώσει :

$$u_i = v_i - p_i$$

Η μελέτη της επιστήμης του σχεδιασμού μηχανισμών αφορά την μελέτη μηχανισμών που ορίζονται όπως παραπάνω και ως σκοπό έχει την μεγιστοποίηση κάποιου μεγέθους. Τα συνηθέστερα μεγέθη είναι είτε η μεγιστοποίηση του αθροίσματος της ωφέλειας των παικτών η οποία ονομάζεται και κοινωνική ωφέλεια $SW = \sum_{i=1}^{n} v_i$, είτε το κέρδος του δημοπράτη $REV = \sum_{i=1}^{n} p_i$. Για τους περισσότερους μηχανισμούς, στόχος είναι να έχουν την ακόλουθη ιδιότητα [27] :

Ορισμός 2.3. ***Φιλαλήθεια***

Ένας μηχανισμός λέμε ότι είναι φιλαλήθης όταν για κάθε παίχτη i η κυρίαρχη (dominant) στρατηγική είναι ο παίχτης να δηλώσει την πραγματική του αξία v_i για το αντικείμενο που τον ενδιαφέρει, ανεξάρτητα από τις κινήσεις των αντίπαλων παιχτών.

2.3 Φιλαλήθεις Δημοπρασίες

2.3.1 Δημοπρασίες First-Price

Ας μελετήσουμε κάποιους βασικούς μηχανισμούς. Στους παραδοσιακούς πλειστηριασμούς γνωρίζουμε τον μηχανισμό σύμφωνα με τον οποίο οι υποψήφιοι υποβάλλουν προσφορές και αυτός με την μεγαλύτερη προσφορά παίρνει το αντικείμενο πληρώνοντας το ποσό της προσφοράς του.

Ο μηχανισμός αυτός αν και γνωστός σε όλους μας δεν είναι φιλαλήθης [10]. Αρκεί να σκεφτούμε ένα απλό παράδειγμα με 2 υποψηφίους. Η αξία του αντικειμένου για τον Α είναι 6 ενώ η αξία του αντικειμένου για τον Β είναι 5. Εάν ο μηχανισμός ήταν φιλαλήθης, έστω ε και δ μικροί αριθμοί, οι παίχτες θα έκαναν προσφορές ίσες με 5+ε και 6+ε ώστε να έχουν κέρδος. Όμως, εάν ο Α γνώριζε την προσφορά του αντίπαλου παίχτη, τότε δεν θα έκανε προσφορά 6+ε αλλά 5+ε+δ ώστε να μεγιστοποιήσει το όφελος του, αφού ενώ πριν το όφελος του ήταν ε, τώρα το όφελος του είναι 6-(5+ε+δ). Επομένως ο παίχτης Α είχε κίνητρο να αλλάξει την προσφορά του ώστε να μην ταυτίζεται με την πραγματική αξία που είχε το αντικείμενο για αυτόν αλλά να την μειώσει ώστε να μεγιστοποιήσει το όφελος του.

Το γεγονός πως ο μηχανισμός δεν είναι φιλαλήθης δημιουργεί προβλήματα στην πρόβλεψη των εσόδων για τον δημοπράτη, ενώ για τους ίδιους τους παίχτες, είναι πολύ δύσκολο να αποφασίσουν με ποιόν τρόπο θα παίξουν.

2.3.2 Δημοπρασίες Second-Price

Ας πάρουμε το ίδιο παράδειγμα και ας θεωρήσουμε πως ο μηχανισμός αλλάζει και ο παίχτης με την μεγαλύτερη προσφορά πληρώνει την τιμή που προσέφερε το άτομο με την 2η μεγαλύτερη προσφορά. Προς υπενθύμιση, για τον Α η αξία του αντικειμένου είναι 6 ενώ για τον Β είναι 5. Έτσι, αν θεωρήσουμε πως οι παίχτες δηλώνουν την τιμή ίση με την αξία που έχει γι΄ αυτούς το αντικείμενο, ας μελετήσουμε ξανά αν έχουν κίνητρο αλλαγής της τιμής τους, αναλυτικά :

- Για τον Α δεν υπάρχει κίνητρο να ανεβάσει την προσφορά του, έστω σε 7, αφού σε περίπτωση που κάποιος άλλος παίκτης ξεπεράσει την αξία που έχει για αυτόν το αντικείμενο (δηλαδή 6), τότε έχουμε 2 πιθανά σενάρια. Είτε ο Α χάνει το αντικείμενο και επομένως έχει όφελος 0, είτε παίρνει το αντικείμενο, αλλά επειδή ο Β θα ξεπεράσει την τιμή 6, ο Α θα έχει όφελος $6 - 7 = -1$ και επομένως θα έχει κόστος και όχι όφελος.

- Για τον Α δεν υπάρχει κίνητρο να κατεβάσει την τιμή αφού η τιμή που θα πληρώσει εάν του κατοχυρωθεί το αντικείμενο είναι η προσφορά του Β και επομένως ανεξάρτητη της προσφοράς του.

- Για τον Β δεν υπάρχει νόημα να ρίξει την τιμή του γιατί και πάλι θα χάσει το αντικείμενο.

- Για τον Β εάν ανεβάσει την προσφορά του και περάσει τον Α, αυτό σημαίνει πως σε περίπτωση που η προσφορά του Α βρίσκεται πάνω απ΄ την αξία του αντικειμένου για τον

Β και επομένως θα έχει και αυτός όφελος αρνητικό και άρα δεν τον συμφέρει να κερδίσει την δημοπρασία.

Επομένως αποδείξαμε πως ο μηχανισμός αυτός είναι φιλαλήθης. Η συγκεκριμένη ιδιότητα είναι εξαιρετικά σημαντική γιατί εάν θεωρήσουμε 2 αντικειμενικούς και έξυπνους παίκτες, μπορούμε να γνωρίζουμε πως έχει εξασφαλιστεί πως οι παίκτες θα δηλώσουν τις πραγματικές τους προτιμήσεις, γεγονός το οποίο θα αποδώσει και μεγαλύτερα κέρδη στον δημοπράτη σε σχέση με τις δημοπρασίες First-Price.

Πέρα απ΄ το παραπάνω, οι δημοπρασίες αυτές έχουν τις ακόλουθες πολύ χρήσιμες ιδιότητες [10] :

- Όλοι οι παίκτες έχουν ωφέλεια μεγαλύτερη ή ίση του μηδενός.
- Οι δημοπρασίες αυτές μεγιστοποιούν την κοινωνική ωφέλεια.
- Η πολυπλοκότητα της δημοπρασίας είναι πολυωνυμική.

2.3.3 Δημοπρασίες Vickrey-Clarke-Groves

Οι δημοπρασίες Vickrey-Clarke-Groves (VCG) είναι μία άλλη προσπάθεια για σχεδιασμό μηχανισμών οι οποίοι είναι φιλαλήθεις. Διαισθητικά, στο VCG ο παίκτης δεν θα χρεωθεί τιμή σχετική με την τιμή που πόνταρε, αλλά θα πληρώσει πόση ζημιά έκανε στους υπόλοιπους παίκτες. Για να υπολογιστεί αυτό, γίνεται σύγκριση της αξίας που λαμβάνει ο κάθε παίχτης στην τωρινή δημοπρασία με την δημοπρασία χωρίς τον εκάστοτε παίχτη i. Η διαφορά της τιμής για τους υπόλοιπους παίχτες (εάν ο i δεν συμμετείχε) αποτελεί και την τιμή που πληρώνει ο i. Με τον τρόπο αυτό, τόσο οι παίκτες αναγκάζονται να κάνουν προσφορές ίσες με την αξία του αντικειμένου γι΄ αυτούς (η απόδειξη βρίσκεται παρακάτω), όσο και μεγιστοποιείται το κοινωνικό όφελος, αφού το αντικείμενο το παίρνουν οι παίκτες για τους οποίους το αντικείμενο αξίζει περισσότερο.

Παράδειγμα

Ας υπολογίσουμε τις τιμές σύμφωνα με το VCG για το ακόλουθο παράδειγμα. Έστω 3 παίκτες και 2 σύνολα. Ο χρήστης μπορεί να αγοράσει μόνο ολόκληρα τα σύνολα κι όχι μέρη τους. Το πρώτο σύνολο αποτελείται από 100 μολύβια ενώ το 2ο αποτελείται από 99 μολύβια. Το κάθε μολύβι έχει για τους 3 παίκτες αξία 10,5 και 1 αντίστοιχα. Ο παίκτης με την αξία 10 ανά μολύβι θα πάρει το πρώτο σύνολο και ο παίκτης με 5, το δεύτερο. Για να υπολογίσουμε τις πληρωμές θα πρέπει να σκεφτούμε τις δημοπρασίες χωρίς τον κάθε παίχτη. Αν ο Α δεν υπήρχε, ο Β θα λάμβανε αξία ίση με $100 \cdot 5 = 500$ ενώ ο Γ $1 \cdot 99 = 99$. Αντίθετα με τον Α στο παιχνίδι, η αξία του Β είναι $99 \cdot 5 = 495$ ενώ του Γ 0. Άρα, ο Α θα πληρώσει $500 + 99 - 495 = 104$ για τα μολύβια 100. Ο Β με τον αντίστοιχο συλλογισμό θα πληρώσει 99.

Όπως παρατηρούμε παραπάνω, οι τιμές στις οποίες καταλήγει ο VCG είναι πολύ χαμηλές, αλλά εξασφαλίζουν πως οι παίχτες θα κάνουν προσφορές ίσες με την πραγματική αξία που έχει για αυτούς το αντικείμενο. Ας ορίσουμε το μοντέλο πιο φορμαλιστικά [22] [27]:

- $A = a_1, a_2, ..., a_m$ αντικείμενα
- $B = b_1, b_2, ..., a_b$ παίχτες
- Έστω V_B^A η κοινωνική ωφέλεια για έναν συγκεκριμένο συνδυασμό αντικειμένων και παιχτών.

Ορισμός 2.1. ***VCG***

Ο παίχτης που θα κερδίσει μία δημοπρασία VCG θα πληρώσει ποσό ίσο με $V_{B\setminus\{b_i\}}^M - V_{B\setminus\{b_i\}}^{M\setminus\{a_j\}}$

Εξηγώντας τον παραπάνω τύπο με λόγια, ο παίχτης που κερδίσει την δημοπρασία θα πληρώσει ποσό ίσο με την διαφορά της κοινωνικής ωφέλειας εάν δεν βρισκόταν ο παίκτης i στο παιχνίδι, μείον την κοινωνική ωφέλεια χωρίς τον παίκτη i και το αντικείμενο j, αφού ο i κέρδισε το αντικείμενο j.

Η απόδειξη της φιλαλήθειας βασίζεται στο ακόλουθο επιχείρημα. Ο παίχτης έχει ωφέλεια $u_i = v_i - p_i$. Αντικαθιστώντας την τιμή που θα πληρώσει για το αντικείμενο, παίρνουμε $u_i = v_i - V_{B\setminus\{b_i\}}^M - V_{B\setminus\{b_i\}}^{M\setminus\{a_j\}}$, όμως ο πρώτος και ο τρίτος όρος αποτελούν το συνολικό κοινωνικό όφελος και επομένως σε έναν VCG μηχανισμό, σκοπός κάθε παίχτη είναι να μεγιστοποιήσει το κοινωνικό όφελος! Με τον τρόπο αυτό, εξασφαλίζεται πως οι παίχτες οφείλουν να ποντάρουν με φιλαλήθη τρόπο και επομένως ο VCG είναι φιλαλήθης μηχανισμός.

2.3.4 Δημοπρασίες Generalized Second-Price

Ορισμός 2.2. ***GSP***

Έστω ότι έχουμε n παίχτες και k αντικείμενα. Θεωρούμε την αξία των αντικειμένων σε αύξουσα σειρά, δηλαδή $v_1 > v_2 > ... > v_k$. Έστω b οι προσφορές των παιχτών. Ταξινομούμε τις προσφορές b. Έστω j η θέση του παίχτη μετά την ταξινόμηση. Εάν $j < k$ ο παίχτης που θα έχει την j-ωστή μεγαλύτερη προσφορά, θα πάρει το αντικείμενο της θέσης j πληρώνοντας ποσό ίσο με $p_j = b_{j+1}$.

Οι περισσότερες δημοπρασίες του σύγχρονου κόσμου, αφορούν περισσότερα του ενός αντικειμένου. Στο πλαίσιο αυτό, το παίγνιο γίνεται πολύ πιο περίπλοκο. Αρχικά αποδεικνύεται [10, 21] ότι ισχύει πως :

Θεώρημα 2.1. *Ο Second-price μηχανισμός για περισσότερα του ενός αντικείμενα δεν είναι φιλαλήθης (GSP).*

Παράδειγμα

Έστω το προηγούμενο παράδειγμα με 3 παίκτες και 2 σύνολα. Ο χρήστης μπορεί να αγοράσει μόνο ολόκληρα τα σύνολα κι όχι μέρη τους. Το πρώτο σύνολο αποτελείται από 100 μολύβια ενώ το 2ο αποτελείται από 99 μολύβια. Το κάθε μολύβι έχει για τους 3 παίκτες αξία 10,5 και 1 αντίστοιχα. Επομένως αν οι 3 παίχτες παίξουν σύμφωνα με την πραγματική αξία που έχουν γι΄ αυτούς τα μολύβια τότε ο παίχτης 1 θα πάρει το πρώτο σύνολο πληρώνοντας 5 (όσο η τιμή της 2ης μεγαλύτερης προσφοράς) για το κάθε μολύβι και επομένως $5 \cdot 100 = 500$. Όμως,

εάν είχε κάνει προσφορά 4.99, θα είχε πάρει το 2ο σύνολο πληρώνοντας $1 \cdot 99 = 99$. Επομένως γίνεται ξεκάθαρο πως δεν είναι η κυρίαρχη στρατηγική οι παίχτες να κάνουν προσφορές όσες και οι πραγματικές τους ανάγκες. Μάλιστα ισχύει πως [10] :

Θεώρημα 2.2. *Εάν οι παίχτες ποντάρουν σύμφωνα με τις πραγματικές του ανάγκες, οι τιμές που διαμορφώνονται από το VCG είναι πάντοτε χαμηλότερες από το GSP.*

Παρ΄ ότι ο μηχανισμός αυτός δεν είναι φιλαλήθης, έχει ορισμένες πολύ χρήσιμες ιδιότητες, για τις οποίες θα μιλήσουμε στο 3ο κεφάλαιο όπου θα δούμε την εφαρμογή του στο πρόβλημά μας.

Κεφάλαιο 3

Επιχορηγούμενες Διαφημίσεις και μοντέλα πρόβλεψης συμπεριφοράς χρήστη

Τα βασικότερα έσοδα των μεγαλύτερων εταιρειών του διαδικτύου όπως η Google και η Yahoo, προέρχονται από δημοπρασίες. Ενδεικτικά, το 2005, από τα 6.14 δισεκατομμύρια δολάρια εσόδων που είχε η Google, πάνω από το 98% προήλθαν από διαφόρων ειδών δημοπρασίες που συνέβησαν από την Google.[10]

Η διπλωματική αυτή μελετά μία συγκεκριμένα κατηγορία δημοπρασιών, αυτές των δημοπρασιών επιχορηγούμενων αναζητήσεων στις μηχανές αναζήτησης. Στο πρόβλημα αυτό ιδιαίτερη σημασία έχει η πρόβλεψη της συμπεριφοράς του χρήστη, στόχος της οποίας είναι η καλύτερη δυνατή πρόβλεψη εάν ο χρήστης θα κάνει κλικ ή όχι, στις εκάστοτε διαφημίσεις.

Στο κεφάλαιο αυτό θα παρουσιαστεί το πλαίσιο των επιχορηγούμενων αναζητήσεων. Αφού γίνει μία εισαγωγή για τους μηχανισμούς με τους οποίους λειτουργεί, θα δούμε πως εφαρμόζονται τα δύο είδη δημοπρασιών που αναφέραμε στο προηγούμενο κεφάλαιο (οι GSP και οι VCG) καθώς και θα γίνει μία σύγκριση μεταξύ τους. Στην συνέχεια, θα παρουσιαστούν τα δύο βασικά μοντέλα της βιβλιογραφίας σχετικά με την πρόβλεψη της συμπεριφοράς του χρήστη, το Αλληλουχίας και το διαχωρίσιμο, ενώ στο τελευταίο υποκεφάλαιο θα παρουσιάσουμε πως συμβαίνει η διαχείριση των διαφημίσεων για τις οποίες δεν έχουμε επαρκεί στατιστικά στοιχεία.

3.1 Δημοπρασίες επιχορηγούμενων αναζητήσεων

3.1.1 Εισαγωγή στο πλαίσιο των διαφημίσεων στις μηχανές αναζήτησης

Η μελέτη της διπλωματικής αυτής, αφορά την μελέτη του προβλήματος των δημοπρασιών που γίνονται για την ανάδειξη των διαφημιζόμενων links στις αναζητήσεις στις μηχανές αναζήτησης στο διαδίκτυο. Το πλαίσιο που μελετάμε λειτουργεί ως εξής:

Ο χρήστης εισάγει στην μηχανή αναζήτησης τις λέξεις ή φράσεις που θέλει να ψάξει. Η μηχανή αναζήτησης εμφανίζει δύο διαφορετικών ειδών προτεινόμενα URLs. Πρώτον, εμφανίζει τις σελίδες τις οποίες αποφασίζει πως είναι πιο σχετικές με το ερώτημα, αποτελέσματα τα οποία θα καλούμε από δω και στο εξής οργανικά. Το άλλο κομμάτι των αποτελεσμάτων αφορά τις σελίδες που αποτελούν διαφημίσεις. Αυτές οι διαφημίσεις είναι χρηματοδοτούμενες από εταιρείες οι οποίες επιθυμούν να διαφημίσουν την ιστοσελίδα της επιχείρησής τους. Τα αποτελέσματα αυτά ονομάζονται επιχορηγούμενα αποτελέσματα (sponsored links) και είναι εμφανώς διαφορετικά από τα οργανικά αποτελέσματα. Αυτό συμβαίνει είτε γιατί εμφανίζονται σε ένα διαφορετικό πλαίσιο πάνω ή δίπλα από τα οργανικά αποτελέσματα, είτε γιατί αναγράφουν δίπλα από το link, επιχορηγούμενο. Η λογική των διαφημίσεων αυτών είναι πως σε διαφορετικές αναζητήσεις, θα εμφανίζονται διαφορετικές χρηματοδοτούμενες διαφημίσεις, καθώς σκοπός είναι οι εταιρείες να στοχεύσουν συγκεκριμένη ομάδα ατόμων. Έτσι για παράδειγμα, εάν αναζητήσουμε "New York Hotels", θα εμφανιστούν στις πρώτες θέσεις της σελίδας διαφημίσεις ξενοδοχείων στην Νέα Υόρκη. Αυτά τα ξενοδοχεία είναι αυτά τα οποία κέρδισαν την δημοπρασία για να διαφημιστούν όταν ένας χρήστης αναζητήσει την φράση "New York Hotels" ή φράσεις κοντινές σε αυτήν.

Όταν ο χρήστης πατήσει την διαφημιζόμενη σελίδα, θα μεταβεί στην σελίδα του διαφημιζόμενου. Αφού ο σκοπός της διαφήμισης επετεύχθει, ο διαφημιζόμενος καλείται να πληρώσει. Υπάρχουν διαφορετικοί τρόποι χρέωσης των διαφημιζόμενων. Αυτό μπορεί να γίνει είτε μόνο για την εμφάνιση της διαφημιζόμενης σελίδας σε μία συγκεκριμένη φράση, είτε να πληρώσουν ανά εμφάνιση σε αναζητήσεις, ή ακόμα και να πληρώσουν ανά αγορά που γίνει από χρήστη ο οποίος μεταβιβάστηκε στην διαφημιζόμενη σελίδα μέσω της μηχανής αναζήτησης. Ο πιο συνηθισμένος μηχανισμός που χρησιμοποιείται ευρέως, είναι η πληρωμή ανά κλικ χρήστη. Έτσι, αρκεί ο χρήστης να κάνει κλικ στην διαφημιζόμενη σελίδα και τότε ο διαφημιζόμενος θα πληρώσει την Google ή την Yahoo ή την εταιρεία στην οποία ανήκει η εκάστοτε μηχανή αναζήτησης, το ποσό που είχαν συμφωνήσει κατά την δημοπρασία. Οι θέσεις που υπάρχουν στην κάθε σελίδα για διαφημιζόμενες σελίδες είναι περιορισμένος και η ψηλότερη θέση στην σελίδα έχει και την ψηλότερη τιμή.

Ένα τυπικό διαφημιστικό σύστημα σε μηχανές αναζήτησης, αποτελείται από τα παρακάτω [28] :

- Ένα σύνολο από διαφημίσεις. Για κάθε διαφήμιση υπάρχει ένας τίτλος, περιγραφή και λέξεις ή φράσεις στις αναζητήσεις των οποίων ο διαφημιζόμενος επιθυμεί να διαφημιστεί. Επίσης κάθε διαφήμιση αντιστοιχίζεται και στην προσφορά που έκανε κατά την δημοπρασία.

- Ένα σύστημα που αντιστοιχίζει την αναζήτηση του χρήστη με τις διαφημίσεις τις οποίες σχετίζονται με την αναζήτηση αυτή.

- Ένα μοντέλο πρόβλεψης των πιθανοτήτων. Το μοντέλο αυτό επιχειρεί να προβλέψει την πιθανότητα να επιλεγεί από τον χρήστη η κάθε μία διαφήμιση.

- Ένα μοντέλο δημοπρασιών. Αυτό το μοντέλο καθορίζει ποιές διαφημίσεις θα διαφημιστούν, καθώς και τα ποσά τα οποία θα πρέπει οι διαφημιζόμενοι να πληρώσουν.

Σύμφωνα με το τελευταίο σημείο, χρειάζεται ένας μηχανισμός ο οποίος θα διαμοιράζει τις θέσεις ανά σελίδα ανάμεσα στους διαφημιζόμενους οι οποίοι παίρνουν μέρος στην δημοπρασία. Οι 2 γνωστοί μηχανισμοί που χρησιμοποιούνται συνήθως στο πλαίσιο των δημοπρασιών είναι ο GSP [10][6][16] και ο VCG [27] που μελετήθηκαν στο προηγούμενο κεφάλαιο.

3.1.2 VCG vs GSP

Ας περιγράψουμε πρώτα πως οι δημοπρασίες που μελετήσαμε στο προηγούμενο κεφάλαιο έχουν εφαρμογή στο πλαίσιο μας. Οι δημοπρασίες στο πλαίσιο μας αφορούν τις θέσεις στις αναζητήσεις των χρηστών στις μηχανές αναζήτησης. Επομένως έχουμε δημοπρασίες πολλαπλών αντικειμένων για τις περιορισμένες θέσεις που υπάρχουν σε κάθε σελίδα εμφάνισης αποτελεσμάτων. Για κάθε φράση ή ομάδα λέξεων, συμβαίνει ξεχωριστή δημοπρασία, ενώ θεωρείται πως η διαφήμιση στην πρώτη θέση έχει μεγαλύτερη αξία από την διαφήμιση στην 2η κ.ο.κ.

Ένα βασικό χαρακτηριστικό που διαχωρίζει το πλαίσιό μας από τα παραδοσιακά πλαίσια, είναι πως ο παίχτης έχει την δυνατότητα να αλλάξει το ποντάρισμά του. Έτσι, οι διαφημιζόμενοι λαμβάνουν μέρος σε ένα συνεχές παίγνιο το οποίο δεν σταματά ποτέ γεγονός που ωθεί τους παίχτες να χρησιμοποιούν ρομπότ τα οποία ποντάρουν συνεχόμενα με σκοπό να μεγιστοποιήσουν το όφελός τους.

Στο πλαίσιο του GSP οι διαφημιζόμενοι δηλώνουν ποιό είναι το μέγιστο ποσό το οποίο είναι διατεθειμένοι να πληρώσουν για κάθε κλικ που λαμβάνουν για μία συγκεκριμένη φράση. Ο διαφημιζόμενος που δήλωσε το μεγαλύτερο ποσό θα λάβει την πρώτη θέση και θα πληρώσει το ποσό που δήλωσε ο διαφημιζόμενος με το 2ο μεγαλύτερο ποσό, + ένα μικρό ποσό. Ο 2ος θα πληρώσει το ποσό του 3ου + ένα μικρό ποσό κ.ο.κ. Αντίθετα, στο VCG ο διαφημιζόμενος θα πληρώσει ποσό ίσο με την αρνητική επιρροή που είχε στους υπόλοιπους διαφημιζόμενους.

Όπως αναφέραμε στο προηγούμενο κεφάλαιο, ο VCG είναι φιλαλήθης και μεγιστοποιεί το κοινωνικό όφελος ενώ ο GSP δεν είναι. Μάλιστα, στον GSP στο πλαίσιό μας, είναι αποδεδειγμένο πως δεν υπάρχει ισορροπία κυρίαρχων στρατηγικών. Παρ΄ όλα αυτά, στην αγορά χρησιμοποιείται ο δεύτερος γεγονός το οποίο με πρώτη ματιά μοιάζει αξιοπερίεργο.

Παρ΄ όλα αυτά, αυτό που πρέπει να λάβουμε υπ΄ όψιν, είναι πως η δημοπρασία είναι δημιούργημα της εταιρείας. Επομένως κίνητρό της δεν είναι να μεγιστοποιήσει το κοινωνικό όφελος αλλά να μεγιστοποιήσει τα δικά της έσοδα. Πιθανοί λόγοι που χρησιμοποιείται ο GSP είναι ότι ο VCG είναι δύσκολο να εξηγηθεί στους διάφορους παίκτες, τα κόστη από την μεταφορά από το ένα πλαίσιο στο άλλο θα είναι μεγάλα , καθώς και πως δεν έχει γίνει πειραματική έρευνα για την συμπεριφορά των χρηστών στο VCG[10]. Επίσης, ενδεχομένως οι τιμές που διαμορφώνονται κατά τον GSP, να σταθεροποιούνται σε επίπεδα μεγαλύτερα των τιμών ισορροπίας κατά το VCG και επομένως μεγαλύτερα έσοδα για τις εταιρείες.

3.2 Μοντέλα Πρόβλεψης Συμπεριφοράς Χρήστη

Η μελέτη της διπλωματικής αυτής επικεντρώνεται κυρίως στην πρόβλεψη της συμπεριφοράς ενός χρήστη. Με σκοπό την μεγιστοποίηση των κερδών κατά τις δημοπρασίες, είναι πολύ σημαντικό να γίνεται ακριβής πρόβλεψη της συμπεριφοράς του χρήστη. Η πρόβλεψη αυτή αφορά την πρόβλεψη των πιθανοτήτων να κάνει κλικ ο χρήστης στην κάθε διαφήμιση που θα προβληθεί σε αυτόν. Τα μοντέλα που χρησιμοποιούνται στην αγορά είναι πολύ απλά ή παραλλαγές αυτών και θα παρουσιαστούν τα βασικότερα εξ αυτών παρακάτω. Για να οριστεί το πρόβλημα πιο φορμαλιστικά, έχουμε :

- n διαφημίσεις
- k θέσεις που προορίζονται για διαφήμιση σε συγκεκριμένη σελίδα
- τα μοντέλα πρόβλεψης επιχειρούν να προβλέψουν τον ρυθμό με τον οποίο επιλέγονται οι εκάστοτε διαφημίσεις (click-through rate, CTR). Το CTR είναι η πιθανότητα ο χρήστης να δει την διαφήμιση και να αποφασίσει να κάνει κλικ σε αυτήν.

Παρακάτω θα παρουσιαστούν τα 2 δημοφιλέστερα μοντέλα πρόβλεψης των CTRs [20, 18, 19, 17, 14].

3.2.1 Διαχωρίσιμο Μοντέλο

Το Διαχωρίσιμο Μοντέλο (Separable) αποτελεί ίσως το απλούστερο μοντέλο. Έστω [17]:

- Έστω a μία συγκεκριμένη διαφήμιση.
- Έστω i μία συγκεκριμένη θέση σε συγκεκριμένη σελίδα.
- Έστω q_a η πιθανότητα της διαφήμισης a να επιλεγεί εάν προβληθεί. Η πιθανότητα αυτή αποτελεί αποτέλεσμα της ποιότητας της διαφήμισης, καθώς και του πόσο σχετική είναι με την φράση που έψαξε ο χρήστης.
- Έστω λ_i η πιθανότητα ο χρήστης να δει την διαφήμιση στην θέση i. Η πιθανότητα αυτή είναι ανεξάρτητη της διαφήμισης και αφορά μόνο την συγκεκριμένη θέση i.

Το Διαχωρίσιμο μοντέλο θεωρεί πως το CTR είναι ίσο με $q_a \cdot \lambda_i$. Το βασικότερο πλεονέκτημα του μοντέλου είναι πως προσφέρει έναν πολύ εύκολο τρόπο για τον δημοπράτη να βρει την βέλτιστη τοποθέτηση των διαφημίσεων. Αυτή γίνεται με την ταξινόμηση των διαφημίσεων κατά το γινόμενο $q_a \cdot b_i$. Με αυτόν τον τρόπο εξασφαλίζεται ότι θα μεγιστοποιηθεί το κέρδος του δημοπράτη. Το επιχείρημα στο οποίο βασίζεται ο ισχυρισμός αυτός αφορά τον τρόπο με τον οποίο γίνεται η χρέωση του διαφημιζόμενου. Για κάθε κλικ που κάνει ένας χρήστης, ο διαφημιζόμενος θα πληρώσει ένα ποσό p_a. Επομένως τα έσοδα προκύπτουν από το γινόμενο του ποσού αυτού p_a επί τον αριθμό των κλικ από χρήστες.

Η συμπεριφορά του χρήστη μοντελοποιείται ως εξής. Ο χρήστης θα εξετάσει τις διαφημίσεις διαδοχικά, επομένως θα ξεκινήσει από την θέση 1, θα συνεχίσει με την θέση 2 κ.ο.κ.

Κάθε θέση την βλέπει με πιθανότητα λ_i, ενώ η πιθανότητα να κάνει κλικ στην εκάστοτε διαφήμιση είναι q_a. Το εάν θα συνεχίσει στην επόμενη θέση, είναι ανεξάρτητο του αν έκανε κλικ στην προηγούμενη διαφήμιση.

3.2.2 Externalities και το Μοντέλο Αλληλουχίας

Το παραπάνω μοντέλο έχει την εμφανή αδυναμία ότι θεωρεί πως η πιθανότητα ο χρήστης να κάνει κλικ σε μία διαφήμιση, είναι ανεξάρτητη από τις υπόλοιπες διαφημίσεις που έχουν εμφανιστεί στην ίδια σελίδα. Έτσι, για παράδειγμα, ενδεχομένως οι χρήστες που κάνουν κλικ στην σελίδα της Wikipedia να μην κάνουν κλικ σε άλλες σελίδες μετά από αυτήν γιατί η σελίδα τους καλύπτει. Έτσι, εάν μία διαφήμιση βρεθεί κάτω από την Wikipedia, που είναι μία πολύ δημοφιλής σελίδα, τότε η πιθανότητα να κάνει κλικ ο χρήστης σε αυτήν, θα έχει μειωθεί σημαντικά.

Για τον λόγο αυτόν, προσπάθησαν να εφαρμοστούν μοντέλα τα οποία λαμβάνουν υπ' όψιν την επιρροή των πιθανοτήτων επιλογής ανάμεσα στις διαφημίσεις. Το δημοφιλέστερο από αυτά αποτελεί το Μοντέλο Αλληλουχίας (Cascade).

Μοντέλο Αλληλουχίας

Το Μοντέλο Αλληλουχίας [20] βασίζεται στην υπόθεση ότι η πιθανότητα μίας διαφήμισης επηρεάζεται από όλες τις προηγούμενες διαφημίσεις. Έτσι, η 3η διαφήμιση επηρεάζεται από την 1η και την 2η κ.ο.κ. Για να εκφραστεί αυτό μαθηματικά ορίζεται το εξής καινούριο μέγεθος [20] :

Ορισμός 3.1. ***Πιθανότητα συνέχειας(Continuation Probability)***

Έστω a μία διαφήμιση. Ως πιθανότητα συνέχειας c_a ορίζουμε την πιθανότητα ο χρήστης να δει την διαφήμιση a και ανεξάρτητα του αν τελικά την επιλέξει, να αποφασίσει να συνεχίσει να εξετάζει τις διαφημίσεις που βρίσκονται κάτω από αυτήν. Αντίστοιχα, με πιθανότητα $1 - c_a$ ο χρήστης σταματά την έρευνα του στην συγκεκριμένη σελίδα.

Ορισμός 3.2. ***Μοντέλο Αλληλουχίας***

Έστω η διαφήμιση a η οποία βρίσκεται στην θέση i. Η πιθανότητα r_{a_i} η διαφήμιση αυτή να επιλεγεί από τον χρήστη είναι :

$r_{a_i} = q_{a_i} \cdot \prod_{j=1}^{i-1} c_{a_j}$

Επομένως, το CTR κατά το Μοντέλο Αλληλουχίας είναι ίσο με την πιθανότητα της διαφήμισης να επιλεγεί εάν προβληθεί, επί το γινόμενο των πιθανοτήτων συνέχειας όλων των προηγούμενων διαφημίσεων. Με αυτόν τον τρόπο, εισήχθηκε η επιρροή της μίας διαφήμισης στην άλλη σε σχέση με το διαχωρισίμο μοντέλο το οποίο θεωρούσε πως η πιθανότητα είναι ανεξάρτητη των γειτονικών διαφημίσεων.

Σε παραλλαγή του μοντέλου Αλληλουχίας, εισάγεται και η επιρροή της θέσης στην πιθανότητα να γίνει κλικ στην διαφήμιση. Έτσι, ο τύπος θα γίνει $r_{a_i} = q_{a_i} \cdot \prod_{j=1}^{i-1} c_{a_j} \cdot \lambda_i$, όπου λ_i η επιρροή της θέσης i στην πιθανότητα να επιλεγεί η διαφήμιση.

Μη Ακολουθιακά κλικ

Σύμφωνα με μία έρευνα των Jeziorski και Segal [19], υπάρχουν ισχυρά στατιστικά στοιχεία τα οποία αποδεικνύουν ότι οι χρήστες δεν λειτουργούν πάντα σύμφωνα με το Μοντέλο Αλληλουχίας. Συγκεκριμένα, σύμφωνα με την έρευνα τους :

- Το 46% των χρηστών που κάνουν κλικ, δεν κάνουν κλικ σε ακολουθιακή σειρά (1,2,3...).
- Το 57% των χρηστών που κάνουν περισσότερα από ένα κλικ, επιλέγουν διαφημίσεις που βρίσκονται σε θέση ψηλότερα από την θέση του προηγούμενού τους κλικ.

Τα παρακάτω αποτελέσματα έρχονται σε αντίθεση με 2 παραπάνω μοντέλα, αφού αυτά υποθέτουν διαδοχικά κλικ. Αποδεικνύεται ότι οι χρήστες λειτουργούν με πολύ πιο περίπλοκο τρόπο που δεν μπορεί να οριστεί φορμαλιστικά με ένα απλό μοντέλο.

Μάλιστα αποδεικνύουν πως μπορεί να υπάρχει είτε θετική είτε αρνητική επιρροή από μία διαφήμιση σε μία άλλη. Έτσι, παρ΄ ότι μία καλή διαφήμιση μπορεί να επηρεάσει τις από κάτω και να μειώσει το CTR τους, μπορεί να συμβεί το αντίθετο και μία κακής ποιότητας διαφήμιση να μεγαλώσει το CTR των υπόλοιπων διαφημίσεων. Παρ΄ ότι η πρώτη επιρροή λαμβάνεται υπ΄ όψιν από το μοντέλο Αλληλουχίας, η δεύτερη κατηγορία δεν προβλέπεται. Επιπλέον αποδεικνύεται ότι υπάρχει ισχυρή επιρροή από τις διαφημίσεις που βρίσκονται σε κατώτερη θέση από την διαφήμιση που εξετάζουμε, γεγονός το οποίο επίσης δεν προβλέπεται από το μοντέλο Αλληλουχίας.

Μία πολύ ενδιαφέρουσα παρατήρηση που έγινε είναι πως εάν ανταλλαχτούν οι διαφημίσεις στις θέσεις 1 και 2, θα αλλάξει το CTR της 3ης διαφήμισης. Το παραπάνω σχετίζεται με τον κορεσμό του χρήστη και επομένως αν βγει μία πολύ κακή διαφήμιση στην πρώτη θέση, ενδεχομένως ο χρήστης να επιλέξει να κάνει καινούρια αναζήτηση.

3.3 Διαχείριση νέων διαφημίσεων

Τα παραπάνω μοντέλα υποθέτουν πως για κάθε διαφήμιση, γνωρίζουμε εκ των προτέρων ορισμένες πιθανότητες που μας είναι χρήσιμες για την πρόβλεψη της συμπεριφοράς του χρήστη. Παρ΄ όλα αυτά, η πραγματικότητα διαφέρει κατά πολύ, αφού οι πιθανότητες αυτές είναι κάτι το οποίο υπολογίζεται εμπειρικά και δεν μπορούμε να γνωρίζουμε από πριν τους αριθμούς αυτούς. Έτσι, αν ζητήσει να διαφημιστεί κάποια ιστοσελίδα για την οποία δεν έχουμε προγενέστερη εμπειρία, οι αριθμοί αυτοί μας είναι άγνωστοι. Μία τέτοια καινούρια διαφήμιση η οποία δεν έχει ξαναδοκιμαστεί ονομάζεται νέα διαφήμιση (cold ad). Μοναδικός τρόπος αποτελεί η δοκιμή των διαφημίσεων σε πραγματικά περιβάλλοντα, με σκοπό την όσο το δυνατόν γρηγορότερη εξερεύνηση (exploration) αυτών των διαφημίσεων και των χαρακτηριστικών τους [4].

Η επιστήμη της εξερεύνησης των διαφημίσεων αποτελεί αρκετά σημαντικό πυλώνα των επιχορηγούμενων αναζητήσεων, αφού πρέπει να βρεθεί η καλύτερη δυνατή ισορροπία ανάμεσα στην μεγιστοποίηση των κερδών ανά σελίδα και στην εξερεύνηση νέων διαφημίσεων. Σε μία νέα διαφήμιση, δεν μπορεί να υπολογιστεί με ακρίβεια το CTR, γι΄ αυτό γίνεται πρόβλεψη του πιθανού CTR. Το γεγονός αυτό μπορεί να δημιουργήσει 2 προβλήματα. Είτε η νέα διαφήμιση

είχε χαμηλότερη ποιότητα από την αναμενόμενη και επομένως προβλήθηκαν διαφημίσεις κακής ποιότητας, είτε αν προβληθεί χαμηλότερη ποιότητα από την πραγματική, διαφημίζονται σελίδες που έχουν χαμηλότερη ποιότητα από αυτές που θα μπορούσαν να διαφημιστούν.

Αντίστοιχα, για τον ίδιο τον διαφημιζόμενο είναι εξίσου σημαντική η ακριβής πρόβλεψη της ποιότητας της διαφήμισης αφού από αυτήν εξαρτάται κατά πολύ η προβολή ή μη της διαφήμισης. Οι βασικοί λόγοι για τους οποίους μία διαφήμιση δεν θα προβληθεί είναι :

- Εάν μία διαφήμιση δεν ξεπεράσει ένα κατώφλι ποιότητας, τότε ανεξάρτητα της χρηματικής προσφοράς που κάνει ο διαφημιζόμενος, τότε η διαφήμιση δεν θα προβληθεί. Για μία μεγάλη εταιρεία όπως η Google, είναι εξαιρετικά σημαντικό να διατηρήσει υψηλή ποιότητα διαφημίσεων ώστε να χτίσει μία φήμη καλών διαφημίσεων, ώστε οι χρήστες να τις εμπιστεύονται μελλοντικά και να τις επιλέγουν.

- Αντίστοιχα υπάρχει ένα κατώφλι του μεγέθους $CTR \cdot bid$ το οποίο κάθε διαφημιζόμενος πρέπει να ξεπεράσει. Εάν ένας διαφημιζόμενος βρεθεί κάτω από το κατώφλι αυτό, δεν θα προβληθεί καμία φορά.

- Τέλος εάν υπάρχει μεγάλος ανταγωνισμός για την συγκεκριμένη φράση ή λέξη, μπορεί ένας διαφημιζόμενος να μην καταφέρει ποτέ να προβληθεί εάν το μέγεθος $CTR \cdot bid$ δεν είναι μεγαλύτερο από τους ανταγωνιστές του. Στο πλαίσιο των δημοπρασιών που βρισκόμαστε, η ποιότητα ισοδυναμεί σε χρήματα αφού οι διαφημιζόμενοι καλούνται να μεγιστοποιήσουν το μέγεθος $CTR \cdot bid$. Έτσι, υψηλότερη ποιότητα σημαίνει χαμηλότερη ανάγκη για ψηλό bid.

Όπως βλέπουμε και οι 3 λόγοι είναι ευαίσθητοι στην κακή πρόβλεψη της ποιότητας από τον μηχανισμό εξερεύνησης.

Multi arm bandit

Σε αυτό το σημείο έχει γίνει ξεκάθαρη η σημασία της εύρεσης της ισορροπίας ανάμεσα στην βέλτιστη τοποθέτηση των διαφημίσεων και στην εξερεύνηση των νέων διαφημίσεων. Στατιστικά, μία νέα διαφήμιση έχει μεγαλύτερη διακύμανση από τις διαφημίσεις για τις οποίες έχουμε μεγάλο αριθμό δεδομένων, γι' αυτό τον λόγο δεν οδηγεί στην καλύτερη δυνατή τοποθέτηση των διαφημίσεων και είναι προτιμότερο να προτιμηθεί μία διαφήμιση με ικανοποιητικό αριθμό εμφανίσεων στο παρελθόν.

Στην βιβλιογραφία, το πρόβλημα που περιγράφουμε ανάγεται στο πρόβλημα multi-arm bandit. Σε αυτό, στον παίχτη δίνονται n χέρια από μηχανές κουλοχέρη και καλείται να παίξει k φορές. Κάθε κουλοχέρης, ακολουθεί κάποια συγκεκριμένη κατανομή η οποία είναι άγνωστη στον παίχτη. Ο παίχτης έχει σκοπό να μεγιστοποιήσει το κέρδος του, το οποίο για να συμβεί απαιτεί μία στρατηγική η οποία θα προσπαθεί ταυτόχρονα τόσο να μεγιστοποιεί το κέρδος, όσο και να εξερευνεί τις κατανομές των διαφορετικών μηχανημάτων. Η αναγωγή στο πρόβλημά μας είναι ξεκάθαρη, με τις μηχανές να αποτελούν τους διαφημιζόμενους, ενώ οι κατανομές να ανάγονται στα $CTRs$ των διαφημίσεων.

Για να παίξει το παιχνίδι ο παίχτης που έχει να λάβει τις αποφάσεις, κάνει την υπόθεση πως κάθε κουλοχέρης έχει μία συγκεκριμένη κατανομή και με κάθε του δοκιμή την ενημερώνει. Έτσι για παράδειγμα, μία συνηθισμένη υπόθεση είναι ότι κάθε κουλοχέρης ακολουθεί την κανονική κατανομή και με κάθε δοκιμή ανανεώνεται η μέση τιμή και η διακύμανση του. Κάποιες συνηθισμένες τεχνικές είναι [28] :

- Δειγματοληψία Thompson : Για κάθε κουλοχέρη υποθέτουμε τις κατανομές που έχουν προκύψει από τις προηγούμενες δοκιμές και διαλέγουμε τον κουλοχέρη με το μεγαλύτερο αναμενόμενο κέρδος.
- UCB : Επιλέγουμε τον κουλοχέρη με το μεγαλύτερο αναμενόμενο κέρδος, συν την απόκλισή του.
- ε-άπληστος : Μόνο σε ε από τους γύρους γίνεται εξερεύνηση ενώ στους υπόλοιπους επιλέγεται ο καλύτερος δυνατός κουλοχέρης.

Με τα παραπάνω ολοκληρώνονται οι γνώσεις που απαιτούνται σχετικές με το πρόβλημα αυτό καθ΄ αυτό. Στο επόμενο κεφάλαιο θα γίνει παρουσίαση των γνώσεων μηχανικής μάθησης που απαιτούνται.

Κεφάλαιο 4

Μηχανική Μάθηση για την πρόβλεψη της συμπεριφοράς του χρήστη

Η πρόβλεψη της συμπεριφοράς του χρήστη μπορεί να συμβεί με χρήση μηχανικής μάθησης. Στο κεφάλαιο αυτό θα γίνει μία εισαγωγή στην μηχανική μάθηση και στην εξόρυξη δεδομένων και θα παρουσιαστεί η λογιστική παλινδρόμηση, αλγόριθμος ο οποίος μπορεί να χρησιμοποιηθεί για την πρόβλεψη της κατηγοριοποίησης σε κλάσεις. Στην συνέχεια παρουσιάζονται λόγοι για τους οποίους θεωρούμε ότι η μηχανική μάθηση θα προσφέρει σημαντικά στην ακρίβεια πρόβλεψης της συμπεριφοράς του χρήστη ενώ τελικά παρουσιάζονται μετρικές αξιολόγησης μοντέλων, τεχνικές οι οποίες χρησιμοποιήθηκαν εκτεταμένα στο 5ο κεφάλαιο που γίνεται η πειραματική αξιολόγηση των μοντέλων της βιβλιογραφίας.

4.1 Εισαγωγή στην Μηχανική Μάθηση και στην εξόρυξη δεδομένων

Η επιστήμη της ανάλυσης δεδομένων και της ανακάλυψης μοτίβων σε αυτά, ονομάζεται εξόρυξη δεδομένων (Data Mining) [23]. Είναι μία αυτόματη ή ημιαυτόματη διαδικασία η οποία εφαρμόζεται σε δεδομένα μεγάλου μεγέθους, με σκοπό την εύρεση συνδυασμών και μοτίβων που ακολουθούν τα δεδομένα αυτά. Αυτή η διαδικασία είναι εξαιρετικά σημαντική αφού την εύρεση αυτών των μοτίβων ακολουθεί η πρόβλεψη μελλοντικών καταστάσεων και αποτελεσμάτων, βγάζοντας μη προφανή συμπεράσματα από τους κανόνες οι οποίοι προέκυψαν από την ανάλυση των δεδομένων.

Στην περίπτωση που αυτά τα μοτίβα οργανωθούν σε δομές οι οποίες μπορούν να εξερευνηθούν και να χρησιμοποιηθούν μελλοντικά για την εξαγωγή συμπερασμάτων, τα μοτίβα αυτά ονομάζονται δομικά. Η μηχανική μάθηση ορίζεται ως η συλλογή από τεχνικές με σκοπό την εύρεση και περιγραφή των δομικών μοτίβων, ένα εργαλείο το οποίο βοηθά στην ερμηνεία των δεδομένων και στην πρόβλεψη των πιθανών αποτελεσμάτων.

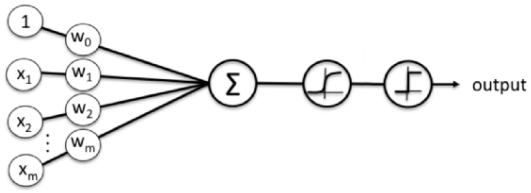
Η μηχανική μάθηση είναι ένας κλάδος που σχετίζεται με την τεχνητή νοημοσύνη και την στατιστική. Απαραίτητο για την μηχανική μάθηση είναι ένα σύνολο δεδομένων τα οποία ονομάζουμε δεδομένα εκπαίδευσης (training data), τα οποία θα χρησιμοποιηθούν για την εκπαίδευση του υπολογιστή ώστε να μπορεί να βγάζει χρήσιμα συμπεράσματα για τα δεδομένα. Η μηχανική μάθηση μπορεί να χωριστεί σε 2 μεγάλες κατηγορίες :

- **Επιβλεπόμενη Μάθηση**, είναι ο κλάδος της μηχανικής μάθησης ο οποίος επιχειρεί να εξάγει μία συνάρτηση η οποία περιγράφει τα δεδομένα που έχουμε στην διάθεσή μας. Κάθε μονάδα δεδομένων αποτελείται από τα χαρακτηριστικά της (features) καθώς και το αποτέλεσμα στο οποίο οδηγεί (label). Ένα παράδειγμα επιβλεπόμενης μάθησης είναι η εκπαίδευση του υπολογιστή να μπορεί να καταλάβει από μία φωτογραφία εάν το ζώο το οποίο εικονίζεται είναι λύκος ή σκύλος, γνωρίζοντας εκ των προτέρων ποιό είναι το ζώο που περιγράφεται στην φωτογραφία για το training set. Αφού η μηχανή εκπαιδευτεί, τότε εάν της δώσουμε μία άγνωστη φωτογραφία, θα είναι σε θέση να βρει ποιό από τα 2 ζώα απεικονίζεται.
- **Μη Επιβλεπόμενη Μάθηση**, αποτελεί την προσπάθεια της εύρεσης μοτίβων στα δεδομένα, για τα οποία όμως δεν γνωρίζουμε τυχόν κατηγοριοποίηση που ακολουθούν (δηλαδή το label είναι άγνωστο). Αντίστοιχο παράδειγμα, θα αποτελούσε αν είχαμε ένα σύνολο φωτογραφιών από ζώα, χωρίς να γνωρίζουμε τι ζώα απεικονίζονται και εμείς να είχαμε ως σκοπό την εύρεση μοτίβων ώστε να ομαδοποιήσουμε τα ζώα που απεικονίζονται στις φωτογραφίες.

Το πρόβλημά μας, αποτελεί πρόβλημα επιβλεπόμενης μάθησης. Ανάλογα με το είδος της εξόδου, μπορούμε να έχουμε και διαφορετικό είδος επιβλεπόμενης μάθησης. Εάν το αποτέλεσμα αποτελεί μία κατηγορία (για παράδειγμα σκύλος ή λύκος), τότε έχουμε ένα πρόβλημα κατηγοριοποίησης (classification). Αντίθετα, αν προσπαθούμε να προβλέψουμε έναν αριθμό, για παράδειγμα μία τιμή ενός προϊόντος, τότε έχουμε ένα πρόβλημα παλινδρόμησης (regression).

4.2 Λογιστική Παλινδρόμηση

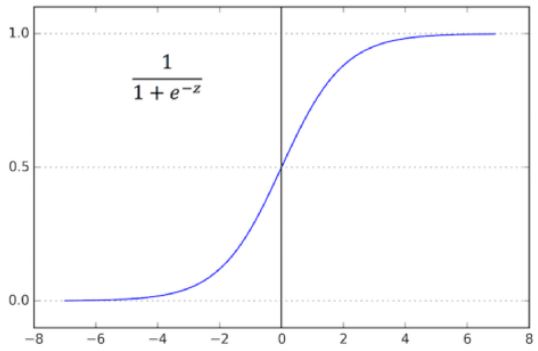
Ένας από τους βασικότερους αλγόριθμους Μηχανικής Μάθησης είναι η λογιστική παλινδρόμηση (logistic regression)[1, 29, 25]. Η λογιστική παλινδρόμηση, παρ΄ ότι περιέχει στο όνομά της την λέξη regression χρησιμοποιείται τόσο για regression όσο και classification. Η λογική της λογιστικής παλινδρόμησης θα εξηγηθεί με την βοήθεια του παρακάτω σχήματος.



Σχήμα 4.1: Logistic Regression

Στο πρώτο στάδιο συμβολίζονται τα χαρακτηριστικά των δεδομένων. Επομένως στο παράδειγμα με τον λύκο και τον σκύλο, πιθανά χαρακτηριστικά μπορεί να είναι το χρώμα, οι διαστάσεις, τα διάφορα χαρακτηριστικά και άλλα. Σκοπός των αλγόριθμων που χρησιμοποιούνται, είναι η εύρεση των κατάλληλων βαρών κάθε μεταβλητής, τα οποία αποτελούν το δεύτερο στάδιο. Με άλλα λόγια, κάθε βάρος εκφράζει το πόσο και πως επηρεάζει το εκάστοτε χαρακτηριστικό το αποτέλεσμα. Στην συνέχεια, τα χαρακτηριστικά με τα βάρη τους συνδυάζονται σε μία συνάρτηση. Η συνάρτηση αυτή είναι σιγμοειδής και έχει ως σκοπό να εξάγει ένα αποτέλεσμα που παίρνει τιμές στο κλειστό διάστημα $[0, 1]$. Στο σημείο αυτό είναι που γίνεται ο διαχωρισμός ανάμεσα σε classification και regression. Εάν επιθυμούμε να κάνουμε regression, τότε μένουμε στο σημείο μετά την έξοδο της σιγμοειδούς συνάρτησης. Το αποτέλεσμα της σιγμοειδούς συνάρτησης αποτελεί μία πιθανότητα η οποία θέλουμε να προβλέψουμε. Αντίθετα, εάν θέλουμε να συνεχίσουμε και να λύσουμε κάποιο πρόβλημα classification θα προχωρήσουμε στο επόμενο βήμα. Σε αυτό, τίθεται ένα κατώφλι, για παράδειγμα 0.5. Οι τιμές που βρίσκονται πάνω απ΄ το 0.5 προβλέπονται ως 1 ενώ αυτές που βρίσκονται κάτω, προβλέπονται ως 0. Επομένως με το τελευταίο βήμα, προσαρμόσαμε το πρόβλημα ώστε να μπορεί να επιλύσει ένα δυαδικό πρόβλημα με εξόδους 0 ή 1.

Πιο αναλυτικά, έστω x_i οι μεταβλητές - χαρακτηριστικά, η πρόβλεψη θα γίνει μέσω της σιγμοειδούς συνάρτησης, παίρνοντας ως έξοδο $h(z) = 1\frac{1}{1+e^{-z}}$, όπου $z = w_0x_0 + w_1x_1 + w_2x_2 + ... + w_nx_n$ και w_i το βάρος του i-οστού χαρακτηριστικού.



Σχήμα 4.2: Σιγμοειδής Συνάρτηση

Σε μορφή διανύσματος, το z μπορεί να γραφτεί ως $z = w^T x$. Δεδομένων των παρακάτω:

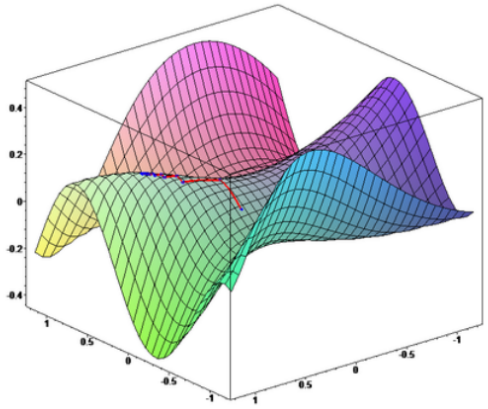
- n εγγραφές (data points).
- m χαρακτηριστικά (features), $(x^0, y^0), (x^1, y^1), ..., (x^{n-1}, y^{n-1})$, όπου x το διάνυσμα των χαρακτηριστικών και y το label

Σκοπός είναι η εύρεση των κατάλληλων βαρών w ώστε να ελαχιστοποιηθεί η παρακάτω συνάρτηση κόστους :

$J(w) = -\frac{1}{n}\sum_{i=0}^{n-1}\{y^i log[h(w^t x^i)] + (1 - y^i) log[1 - h(w^T x^i)]\}$

Ένας απ΄ τους συνηθέστερους αλγόριθμους εύρεσης του ελαχίστου της συνάρτησης κόστους είναι ο Gradient Descent. Ο αλγόριθμος αυτός ξεκινά από ένα σύνολο τιμών w και επανα-

ληπτικά κινείται προς την εύρεση του ελαχίστου της συνάρτησης κόστους $J(w)$. Αυτό επιτυγχάνεται με την κίνηση προς την αντίθετη κατεύθυνση από την παράγωγο της συνάρτησης κόστους. Πιο απλά, ο αλγόριθμος ακολουθεί την κατεύθυνση της κλίσης που δημιουργεί η επιφάνεια της συνάρτησης κόστους, μέχρι να βρεθεί σε κοιλάδα. Το παρακάτω σχήμα παρουσιάζει σχηματικά τα παραπάνω.



Σχήμα 4.3: Gradient Descent

Επομένως για κάθε επανάληψη, το βάρος ανανεώνεται προς την αντίθετη κατεύθυνση της παραγώγου, με ρυθμό εκμάθησης(learning rate) α :

$\delta w = -a\nabla_w J(w)$

4.3 Πρόβλεψη συμπεριφοράς του Χρήστη με χρήση Μηχανικής Μάθησης

Η μηχανική μάθηση στο πλαίσιό μας, μπορεί να χρησιμοποιηθεί για την πρόβλεψη της συμπεριφοράς του χρήστη με μεγαλύτερη ακρίβεια. Συγκεκριμένα θα παρουσιαστούν ορισμένα παραδείγματα σεναρίων που δεν λαμβάνονται υπ΄ όψιν από τα μαθηματικά μοντέλα που παρουσιάστηκαν παραπάνω και επομένως τεχνικές μηχανικής μάθησης θα βοηθήσουν στην αύξηση της ακρίβειας με την οποία προβλέπεται η συμπεριφορά του χρήστη.

Σενάρια Υπεροχής Μηχανικής Μάθησης

- Εάν ο χρήστης αναζητήσει το όνομα ενός συγκεκριμένου ιστότοπου, τότε οι πιθανότητες να επιλέξει ο κάθε χρήστης τις διαφημίσεις αλλάζουν δραματικά. Αυτό συμβαίνει καθώς ο χρήστης που αναζήτησε τον ιστότοπο a, θα επιλέξει την διαφήμιση του ιστότοπου a με πολύ μεγάλη πιθανότητα ενώ οι πιθανότητες να επιλέξει τις υπόλοιπες διαφημίσεις θα είναι σχεδόν μηδενικές. Πρόκειται για ένα σενάριο που δεν μπορεί να ληφθεί υπ΄ όψιν από τα απλά μοντέλα όπως το Αλληλουχίας και το διαχωρίσιμο.

- Αντίστοιχα, η αναζήτηση ενός συγκεκριμένου προϊόντος που πωλείται από μία μόνο

σελίδα μπορεί να οδηγήσει έναν χρήστη στην αντίστοιχη συμπεριφορά με την παραπάνω περίπτωση.

- Τα παραπάνω σενάρια μπορεί να γενικευτούν απ΄ τον εξής κανόνα: ο κάθε ιστότοπος μπορεί να έχει διαφορετικά $CTRs$ ανάλογα με την λέξη ή φράση της αναζήτησης που κάνουν. Έτσι, κάποιο ιστότοπος μπορεί να είναι πολύ δημοφιλής για ποδόσφαιρο ενώ είναι λιγότερο για μπάσκετ, ενώ ο ιστότοπος Wikipedia είναι κυρίαρχη διαφήμιση όποτε γίνεται αναζήτηση της λέξης wiki.

- Ένα άλλο χαρακτηριστικό είναι πως ανάλογα την λέξη, μπορεί να διαφέρει ο αριθμός των κλικ που κάνει ένας χρήστης. Για παράδειγμα, εάν γίνει μία αναζήτηση με την φράση ¨καιρός Αθήνα¨, είναι πολύ πιθανό πως ο χρήστης θα κάνει μόνο ένα κλικ, καθώς είναι σίγουρος πως οποιονδήποτε ιστότοπο και να διαλέξει, θα πάρει το αποτέλεσμα το οποίο θέλει να αναζητήσει.

- Αντίθετα, εάν η αναζήτηση είναι ¨παραγγελία παπούτσια¨, ο χρήστης σημαίνει πως ενδιαφέρεται να αγοράσει παπούτσια. Αυτό σημαίνει πως είναι πολύ πιο πιθανό να επιλέξει πολλούς ιστότοπους, καθώς θα ενδιαφερθεί να κάνει έρευνα αγοράς για να βρει την κατάλληλη προσφορά. Στην περίπτωση αυτή, οι πιθανότητες αυξάνονται όταν η έρευνα είναι σχετική με αγορά από τον χρήστη.

- Μία άλλη περίπτωση αφορά τις λέξεις που σχετίζονται με αναζήτηση φωτογραφιών και όχι ιστότοπων. Έτσι, εάν ο χρήστης κάνει αναζήτηση η οποία υπονοεί ότι ο χρήστης θέλει να βρει φωτογραφίες. Έτσι και πάλι, οι πιθανότητες των διαφημίσεων πέφτουν κατακόρυφα, αφού ο χρήστης δεν ενδιαφέρεται για ιστότοπο αλλά για φωτογραφίες.

- Τελευταίο αλλά και πολύ σημαντικό αποτελεί η ανάλυση της συμπεριφοράς του κάθε χρήστη μεμονωμένα. Αυτό σημαίνει πως διαφορετικοί χρήστες έχουν διαφορετικές συμπεριφορές στις ίδιες καταστάσεις (έρευνα των Athey, Ellison (2011) [3]). Επομένως μπορεί να υπάρχει χρήστης ο οποίος διαλέγει το πρώτο αποτέλεσμα πάντοτε, χρήστης ο οποίος δεν διαλέγει ποτέ διαφημίσεις ή χρήστης ο οποίος επιλέγει να ανοίξει πάντοτε τα περισσότερα από τα αποτελέσματα τα οποία προκύπτουν. Επομένως πιο ακριβής ανάλυση μπορεί να πραγματοποιηθεί μέσω της ανάλυσης της συμπεριφοράς κάθε χρήστη μεμονωμένα.

Τα παραπάνω αποτελούν λόγους για τους οποίους μπορεί να προτιμηθεί η ανάλυση μέσω μηχανικής μάθησης έναντι της στατιστικής ανάλυσης. Η ανάλυση αυτή προϋποθέτει την ύπαρξη μεγάλου όγκου δεδομένων τα οποία παρουσιάζουν αναλυτικά την συμπεριφορά του χρήστη και δεν αποκρύπτουν δεδομένα. Δυστυχώς, η εύρεση τέτοιων δεδομένων είναι ανέφικτη καθώς κάθε εταιρεία κρατά τα δεδομένα της αποκλειστικά για δική της ανάλυση και χρήση τόσο για λόγους ανταγωνισμού όσο και για λόγους προστασίας των δεδομένων (στο 5ο κεφάλαιο γίνεται μία σύντομη αναφορά στις επιπτώσει που είχε για την AOL η δημοσίευση κάποιων φαινομενικά απρόσωπων δεδομένων). Για τον λόγο αυτό η διπλωματική ολοκληρώθηκε με

στατιστική ανάλυση, προς την κατεύθυνση της εύρεσης καλύτερων μοντέλων από το διαχωρίσιμο και το Αλληλουχίας, τα οποία επιτυγχάνουν πιο ακριβή πρόβλεψη της συμπεριφοράς του χρήστη.

4.4 Βασικές Μετρικές Αξιολόγησης Regression

Μετά την εκπαίδευση ενός μοντέλου μηχανικής μάθησης, ακολουθεί η αξιολόγηση του. Στην μηχανική μάθηση η εκπαίδευση συμβαίνει με τα δεδομένα εκπαίδευσης (training set). Η αξιολόγησή του γίνεται με την σύγκριση των προβλεπόμενων τιμών, με τις πραγματικές. Όμως, επειδή η εκπαίδευση έχει γίνει με χρήση του training set, θα ήταν λάθος η αξιολόγηση να γίνει πάνω στα ίδια δεδομένα. Για τον λόγο αυτό, χρειάζεται ένα άλλο σύνολο δεδομένων (test set), στο οποίο θα γίνει η αξιολόγηση. Ανάλογα με την επιλογή μας ανάμεσα σε regression και classification, έχουμε διαφορετικές μετρικές αξιολόγησης του μοντέλου που έχουμε εκπαιδεύσει. Παρακάτω θα παρουσιαστούν οι πιο βασικές μετρικές αξιολόγησης για regression. Για να οριστούν αυτές, ας ορίσουμε πρώτα το πλαίσιο στο οποίο βρισκόμαστε [32] :

- Έχουμε n εγγραφές
- Η προβλεπόμενη τιμή $\hat{y}_i$ της i-οστής εγγραφής.
- Η πραγματική τιμή y_i της i-οστής εγγραφής.
- **Μέσο Απόλυτο Σφάλμα - Mean Absolute Error - MAE**. $\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$

  Υπολογίζεται από την απόλυτη διαφορά ανάμεσα στην πρόβλεψη του μοντέλου και την πραγματική τιμή. Οι διαφορές αυτές αθροίζονται και διαιρούνται με τον αριθμό των τιμών ώστε να βρεθεί ο μέσος όρος αυτής της ποσότητας. Είναι ίσως η πιο βασική μετρική η οποία μπορεί να δείξει ποσοτικά την απόσταση του μοντέλου από την πραγματικότητα. Παρ΄ ότι επιτυγχάνει να απεικονίσει την ποσότητα της διαφοράς, δεν μπορεί να δείξει την κατεύθυνση της διαφοράς. Επομένως δεν μπορούμε να γνωρίζουμε αν το μοντέλο προβλέπει τιμές μεγαλύτερες ή μικρότερες από τις πραγματικές.

- **Μέσο Ριζικό Τετραγωνικό Σφάλμα - Root Mean Square Error - RMSE**. $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$

  Πρόκειται για την ρίζα του μέσου όρου της τετραγωνικής διαφορά πρόβλεψης - πραγματικής τιμής. Είναι μία μετρική πολύ κοντά στην προηγούμενη, με την διαφορά ότι οι διαφορές υψώνονται στο τετράγωνο και στην συνέχεια παίρνουμε την ρίζα του συνολικού μεγέθους. Έχει παρόμοια χαρακτηριστικά με το MAE με βασική διαφορά πως λόγω του τετραγώνου, τιμωρεί περισσότερο τις μεγαλύτερες αποκλίσεις. Έτσι, εάν υπάρχουν τιμές που απέχουν πολύ από τις πραγματικές, τότε θα έχουν μεγαλύτερο βάρος σε σχέση με τις τιμές κοντά στις πραγματικές και επομένως θα επηρεάσουν σημαντικά στον υπολογισμό του RMSE.

- **R^2**

  Για να οριστεί η μετρική R^2 χρειάζεται πρώτα να ορίσουμε τα ακόλουθα μεγέθη :

  - $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
  - $SS_{tot} = \sum_{i=1}^{n}(y_i - \overline{y})^2$
  - $SS_{res} = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$

  Τελικά το R^2 ορίζεται ως $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$

  Όπως και οι 2 παραπάνω μετρικές, το R^2 δείχνει πόσο καλά το μοντέλο προβλέπει. Το R^2 όμως έχει 2 πολύ χρήσιμες ιδιότητες. Η πρώτη είναι ότι παίρνεις τιμές στο διάστημα $[0, 1]$. Η δεύτερη είναι ότι αποτελεί μία μετρική συγκριτική και όχι απόλυτη. Το R^2 δείχνει πόσο καλύτερα ο προβλέπτης μας αποδίδει σε σχέση με έναν προβλέπτη ο οποίος προβλέπει σύμφωνα με τον μέσο όρο των τιμών. Επομένως όσο μεγαλύτερη τιμή έχουμε, τόσο καλύτερα το μοντέλο μας περιγράφει τα δεδομένα.

4.5 Μετρικές Classification

4.5.1 Πίνακας σύγχυσης

Ο πίνακας σύγχυσης (confusion matrix) [33] είναι μία απλή τεχνική που χρησιμοποιείται σε classification προβλήματα. Έστω ότι έχουμε 2 κλάσεις την θετική και την αρνητική. Σε αυτόν παρουσιάζονται τα ακόλουθα :

- Ποιές εγγραφές δεδομένων προβλέφθηκαν σωστά ως θετικές **True Positive-TP**
- Ποιές εγγραφές δεδομένων προβλέφθηκαν σωστά ως αρνητικές **True Negative-TN**
- Ποιές εγγραφές δεδομένων προβλέφθηκαν λανθασμένα ως θετικές **False Positive-FP**
- Ποιές εγγραφές δεδομένων προβλέφθηκαν λανθασμένα ως αρνητικές **False Negative-FN**

Ο παρακάτω πίνακας βοηθά στο να αποκτήσουμε μία πιο συγκεκριμένη εικόνα για το που ο προβλέπτης προβλέπει σωστά και που κάνει λάθος. Τελικά η ακρίβεια (accuracy) ορίζεται ως :

Ορισμός 4.1. ***Ακρίβεια***

$$\frac{TP+TN}{TP+TN+FN+FP}$$

4.5.2 ROC Curve και AUC

Όπως αναφέρθηκε και στο κεφάλαιο του logistic regression, regression και classification μπορούν να βρίσκονται πολύ κοντά και επομένως το ένα πρόβλημα να αναχθεί στο άλλο. Μία πολύ σημαντική μετρική που χρησιμοποιείται στην διπλωματική αυτή είναι οι γραφικές

παραστάσεις ROC. Η διαδικασία δημιουργίας της γραφικής παράστασης είναι η εξής [26] [9] [12]:

- Όπως αναφέραμε, η μετάβαση από regression σε classification χρειάζεται ο ορισμός ενός κατωφλίου απ΄ το οποίο και πάνω θεωρείται πως η εγγραφή ανήκει στην κλάση 1, ενώ κάτω από αυτό ανήκει στην τιμή 0. Για την δημιουργία της γραφικής παράστασης, χρειάζεται ένα μεταβλητό κατώφλι. Οι τιμές που παίρνει στην διπλωματική αυτή είναι από 0 έως 1 με βήμα 0.1.

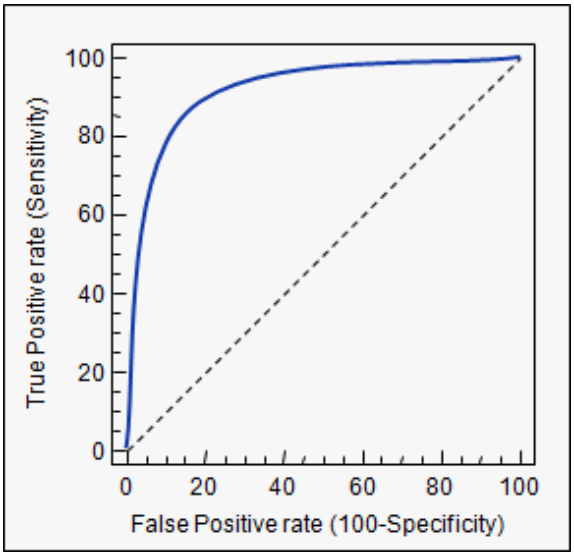
- Για κάθε κατώφλι θα υπολογιστούν ξεχωριστά ποιές εγγραφές κατηγοριοποιήθηκαν στην κλάση 0 και ποιες στην κλάση 1.

- Επίσης για κάθε κατώφλι πρέπει να υπολογιστούν τα ακόλουθα μεγέθη :
  TP : τις τιμές που σωστά προβλέφθηκαν ως μέρος της κλάσης 1
  TN : τις τιμές που σωστά προβλέφθηκαν ως μέρος της κλάσης 0
  FP : τις τιμές που λανθασμένα προβλέφθηκαν ως μέρος της κλάσης 1
  FN : τις τιμές που λανθασμένα προβλέφθηκαν ως μέρος της κλάσης 0

- Στην συνέχεια υπολογίζονται τα ακόλουθα μεγέθη :
  sensitivity $= \frac{TP}{TP+FN}$, το οποίο αποτελεί τον λόγο σωστών προβλέψεων ως 1 δια των συνολικών προβλέψεων ως 1.
  specificity $= \frac{TN}{TN+FP}$, το οποίο αποτελεί τον λόγο σωστών προβλέψεων ως 0 δια των συνολικών προβλέψεων ως 0.

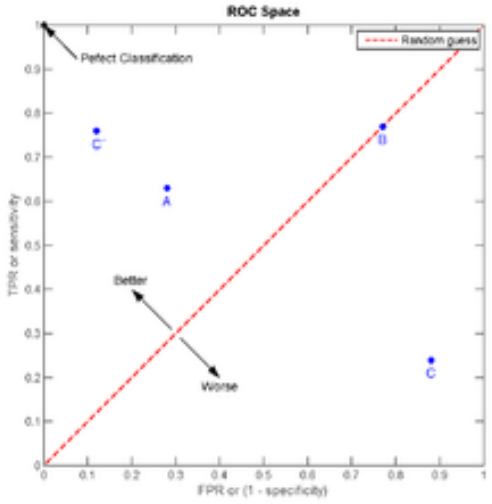
- Η γραφική παράσταση δημιουργείται από τα παραπάνω, με τον άξονα y να αποτελεί το sensitivity και τον άξονα x να αποτελεί το $1 - specificity$.

Ενδεικτικά παρακάτω παρουσιάζεται μία γραφική παράσταση :



Σχήμα 4.4: ROC curve

Η μπλε γραμμή αποτελεί την γραφικά παράσταση ROC ενός προβλέπτη ενώ με διακεκομμένη ευθεία παρουσιάζεται ο τυχαίος προβλέπτης, ο οποίος βρίσκεται στην ευθεία $x = y$. Όσο καλύτερο προβλέπτη έχουμε, τόσο μεγαλύτερο εμβαδό δημιουργείται ανάμεσα στην καμπύλη και την ευθεία $x = 0$. Επομένως όσο πιο πάνω βρίσκεται η ROC ενός προβλέπτη έναντι της $x = y$, τόσο πιο πετυχημένος είναι ο προβλέπτης. Σχηματικά αυτό φαίνεται παρακάτω :



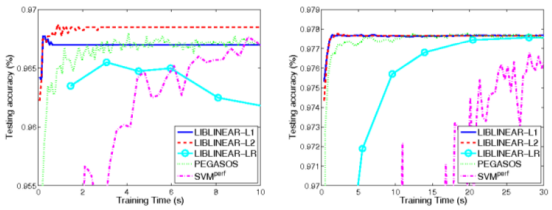
Σχήμα 4.5: ROC curve

Για να ποσοτικοποιηθεί το πόσο καλός είναι ο προβλέπτης λαμβάνουμε το εμβαδό κάτω από την ROC. Το εμβαδό αυτό ονομάζεται Area Under the Curve - AUC. Αυτό συμβαίνει καθώς λόγω του μεταβλητού κατωφλίου, λαμβάνουμε όχι μόνο πόσες σωστές προβλέψεις έχουμε αλλά και πόσο κοντά βρίσκονται οι προβλέψεις στην σωστή τιμή.

4.6 Εισαγωγή στο εργαλείο LIBLINEAR

Το LIBLINEAR [11] είναι μία open source βιβλιοθήκη η οποία δημιουργήθηκε από μέλη του Computer Science Department του Πανεπιστημίου της Ταϊβάν. Πρόκειται για ένα εργαλείο το οποίο υποστηρίζει logistic regression και linear support vector machines και είναι ιδιαίτερα αποδοτικό σε μεγάλο όγκο δεδομένων.

Ενδεικτικά, σε συγκρίσεις του liblinear με άλλα εργαλεία όπως το Pegasos ή το SVM^{perf} στα ίδια δεδομένα, το liblinear φάνηκε να υπερτερεί τόσο σε ταχύτητα όσο και σε ακρίβεια. Οι παρακάτω γραφικές παραστάσεις δίνονται από τους ίδιους τους δημιουργούς του liblinear οπότε παρατίθενται εδώ με επιφυλάξεις :



Σχήμα 4.6: Σύγκριση LIBLINEAR με άλλα εργαλεία

Περισσότερα για την εσωτερική λειτουργία του liblinear βρίσκονται στο paper των δημιουργών του εργαλείου αυτού. Μπορεί να χρησιμοποιηθεί είτε ξεχωριστά από terminal, είτε ως βιβλιοθήκη σε python, C++,Matlab. Διαθέτει 7 διαφορετικούς solvers για classification και 3 για regression, ενώ δίνει την ευκαιρία στον χρήστη να ασχοληθεί με την βελτιστοποίηση των παραμέτρων κάθε διαφορετικού solver.

Κεφάλαιο 5

Ανάλυση δεδομένων και πειραματική αξιολόγηση δεδομένων

Στο κεφάλαιο αυτό, έχοντας ολοκληρώσει την παρουσίαση των θεωρητικών γνώσεων που είναι απαραίτητες για την κατανόηση του προβλήματός μας, θα γίνει η πειραματική αξιολόγηση των μοντέλων της βιβλιογραφίας. Θα παρουσιαστούν τα δεδομένα που χρησιμοποιήθηκαν, θα γίνει μία εισαγωγή στο Hadoop το οποίο αποτελεί το εργαλείο που χρησιμοποιήθηκε για την επεξεργασία των δεδομένων λόγω του μεγάλου μεγέθους τους, ενώ θα γίνει αναλυτική περιγραφή της διαδικασίας επεξεργασίας και ανάλυσης των δεδομένων. Τελικά θα συγκριθούν τα μοντέλα της βιβλιογραφίας με τις μετρικές που παρουσιάστηκαν στο 4ο κεφάλαιο, ενώ θα προταθούν και εναλλακτικά μοντέλα, τα οποία βρίσκονται κοντά σε ακρίβεια σε σχέση με τα μοντέλα της βιβλιογραφίας.

5.1 Περιγραφή Δεδομένων

Για την μελέτη του προβλήματος της πρόβλεψης της συμπεριφοράς του χρήστη, το πρώτο βήμα ήταν η εύρεση δεδομένων. Καθώς η εύρεση τέτοιων δεδομένων φάνηκε ακατόρθωτη, έγινε η υπόθεση πως η συμπεριφορά των χρηστών στα οργανικά αποτελέσματα (τα φυσικά αποτελέσματα που επιλέγονται μόνο λόγω της σχετικότητας τους και δεν προκύπτουν από δημοπρασίες) είναι η ίδια με την συμπεριφορά τους στα χρηματοδοτούμενα αποτελέσματα. Πρόκειται για μία υπόθεση που έχει ξαναειπωθεί και προς το παρόν δεν υπάρχει απόδειξη ούτε υπέρ ούτε κατά της. Πρώτα θα παρουσιαστούν δύο σύνολα δεδομένων (datasets) τα οποία βρέθηκαν αλλά τελικά αποφασίστηκε ότι οι πληροφορία που περιείχαν δεν ήταν αρκετή.

5.1.1 Εναλλακτικά Datasets

Παρακάτω θα παρουσιαστούν δύο Datasets που βρέθηκαν καθώς και οι λόγοι για τους οποίους αποφασίστηκαν ότι δεν είναι ικανοποιητικά.

A3 - Yahoo! Search Marketing Advertiser Bid-Impression-Click data on competing Keywords

Το συγκεκριμένο Dataset προσφέρεται ελεύθερα στο διαδίκτυο από την Yahoo για ερευνητικούς σκοπούς. Πρόκειται για δεδομένα τα οποία εστιάζουν στους διαφημιζόμενους και στις προσφορές που κάνουν για να κερδίσουν τους πλειστηριασμούς. Είναι το μόνο σύνολο δεδομένων το οποίο αποτελείται από δεδομένα επιχορηγούμενων αναζητήσεων. Τα στοιχεία που περιέχει είναι :

- Την ημέρα.
- Το ID του διαφημιζόμενου.
- Τις λέξεις που ενδιαφέρουν τον διαφημιζόμενο να διαφημιστεί (σε κωδικοποιημένη μορφή ώστε να μην δίνει πληροφορία σχετικά με το ποιά λέξη ήταν).
- Την θέση στην οποία τελικά διαφημίστηκε η διαφήμιση.
- Το μέσο ποντάρισμα που έκανε ο συγκεκριμένος διαφημιζόμενος.
- Τον αριθμό των εμφανίσεων της διαφήμισης.
- Τον αριθμό των κλικ.

Ο βασικός λόγος για τον οποίο δεν χρησιμοποιήθηκαν τα δεδομένα αυτά είναι πως εστιάζουν στην πλευρά του διαφημιζόμενου, καθώς και πως δεν είναι εφικτό να αναπαράγουμε τις σελίδες στις οποίες προβλήθηκε. Με τον τρόπο αυτό δεν μπορούμε να μελετήσουμε την συμπεριφορά του χρήστη καθώς δεν γνωρίζουμε ανάμεσα σε ποιές διαφημίσεις είχε να επιλέξει. Τα δεδομένα αυτά θα είχαν χρήση για έρευνα σχετικά με τις δημοπρασίες και τις τιμές που διαμορφώνονται σε αυτές.

AOL Data Leak

Τα δεδομένα αυτά δημοσιεύτηκαν στο διαδίκτυο από την AOL για ερευνητικούς σκοπούς. Δυστυχώς γι΄ αυτήν, δημοσίευσε τα δεδομένα χωρίς καμία επεξεργασία, γεγονός το οποίο οδήγησε σε ταυτοποίηση των χρηστών μετά από έρευνες γεγονός το οποίο οδήγησε σε νομικές περιπέτειες την εταιρεία και σε παραίτηση μελών των ερευνητών της εταιρείας. Τα δεδομένα αυτά παρουσιάζουν τις συνεδρίες (sessions) διαφόρων χρηστών με τις λέξεις που αναζήτησαν, καθώς και το αποτέλεσμα που επέλεξαν. Πιο συγκεκριμένα :

- Το ID του χρήστη.
- Την φράση που αναζήτησε.
- Την ώρα που έκανε την έρευνα.
- Την θέση στην οποία βρισκόταν η σελίδα που επέλεξε ο χρήστης.
- Το URL της σελίδας που επέλεξε.

Παρ΄ ότι η πληροφορία σχετικά με τις λέξεις θα ήταν πολύ βοηθητική, από τα δεδομένα αυτά λείπει η πληροφορία σχετικά με τις ιστοσελίδες οι οποίες εμφανίστηκαν μαζί με την σελίδα την οποία επέλεξε ο χρήστης. Έτσι και σε αυτό το Dataset δεν μπορούμε να γνωρίζουμε ποιές σελίδες απέρριψε ο χρήστης αλλά μόνο ποιές επέλεξε να ανοίξει.

5.1.2 Yandex Personalized Web Search

Τα δεδομένα που τελικά χρησιμοποιήθηκαν δόθηκαν από την Yandex στα πλαίσια ενός διαγωνισμού οργανωμένου σε συνεργασία της Yandex με την πλατφόρμα Kaggle. Η Yandex αποτελεί μία ρώσικη εταιρεία διαδικτυακών υπηρεσιών, η οποία διαθέτει και μηχανή αναζήτησης που επιστρέφει ρώσικα αποτελέσματα. Όλα τα δεδομένα έχουν γίνει ανώνυμα ενώ λέξεις και URLs έχουν κωδικοποιηθεί σε μορφή αριθμητική ώστε ο χρήστης να μην μπορεί να βγάλει συμπεράσματα για τις αναζητήσεις. Έχουν μέγεθος 16GB και οι πληροφορίες που περιέχουν είναι οι ακόλουθες :

- Session ID.
- Την ημέρα.
- Το ID του χρήστη.
- Τον χρόνο που μεσολάβησε ανάμεσα στην αναζήτηση και την επιλογή του URL.
- Το ID του ερωτήματος που αναζήτησε ο χρήστης.
- Την λίστα από τις λέξεις που αναζήτησε ο χρήστης, όλες σε κωδικοποιημένη μορφή.
- Το ID της σελίδας αποτελεσμάτων.
- Το ID του URL που επιλέχθηκε.
- Την λίστα από τα URLs και τα domains των αποτελεσμάτων που εμφανίστηκαν, όλα σε μορφή ID.

Να αναφερθεί πως όταν αναφέρουμε URL εννοείται η συγκεκριμένη σελίδα ενός ολόκληρου ιστότοπου ενώ ως domain αναφέρετε ο ιστότοπος συνολικά. Επομένως Domain είναι η Wikipedia ενώ URL είναι η σελίδα της Wikipedia για την Θεωρία Παιγνίων για παράδειγμα. Τα δεδομένα αυτά έχουν την πληροφορία των αποτελεσμάτων της αναζήτησης και γι΄ αυτό αποφασίστηκε πως θα γίνει μελέτη σε αυτά. Δυστυχώς οι αναζητήσεις ήταν κωδικοποιημένες και επομένως δεν ήταν εφικτό να γίνει μελέτη με τεχνικές μηχανικής μάθησης. Η ακριβής δομή των δεδομένων και η επεξεργασία που έγινε σε αυτά θα γίνει στο μεθεπόμενο υποκεφάλαιο.

5.2 MapReduce

Το MapReduce είναι ένα προγραμματιστικό μοντέλο παράλληλης επεξεργασίας μεγάλου όγκου δεδομένων (Big Data). Στο υποκεφάλαιο αυτό θα παρουσιαστεί συνοπτικά η χρησι-

μότητα του μοντέλου αυτού, ενώ στην συνέχεια θα παρουσιαστεί η πιο διαδεδομένη υλοποίηση του MapReduce [7] [15] το Hadoop [31].

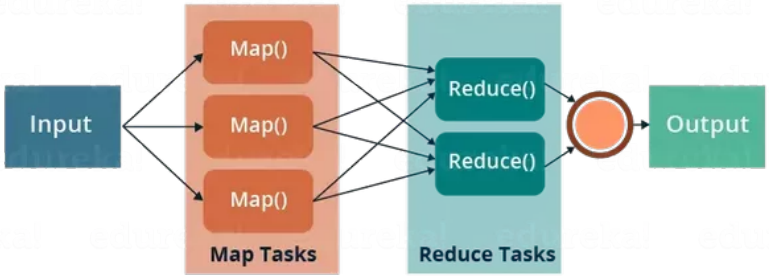
5.2.1 MapReduce Framework

Το MapReduce Framework αναλαμβάνει μία σειρά από πολύ σημαντικές λειτουργίες.

- Την διαμέριση δεδομένων στο cluster.
- Την διαχείριση των παράλληλων εργασιών.
- Την διαχείριση αποτυχιών.
- Την εξισορρόπηση των δεδομένων στους επεξεργαστές
- Την επικοινωνία ανάμεσα στους κόμβους του cluster.

Με τον τρόπο αυτό, ο χρήστης μπορεί να εμπιστευθεί όλα τα παραπάνω στο Framework, αρκεί να προγραμματίσει με την προγραμματιστική λογική του map-reduce. Η λογική του MapReduce απαιτεί σε κάθε κώδικα να είναι ορισμένες δύο συναρτήσεις, η map και η reduce. Η map λαμβάνει ως είσοδο ένα ζευγάρι $(key, value)$, τα επεξεργάζεται και παράγει μία έξοδο της ίδιας μορφής. Στην συνέχεια, τα ζευγάρια αυτά ομαδοποιούνται σύμφωνα με το κλειδί και δίνονται ως είσοδος στην reduce. Με τον τρόπο αυτό η είσοδος της reduce είναι (key , list of values). Το αποτέλεσμα της reduce είναι και τελικά το αποτέλεσμα του προγράμματος.

Τα παραπάνω φαίνονται συγκεντρωμένα στο παρακάτω σχήμα :



Σχήμα 5.1: MapReduce

Επομένως :

$Map: : (key_1, value_1) \rightarrow list(key_2, value_2)$

Η Map εφαρμόζεται παράλληλα στα κομμάτια που έχει διαχωριστεί η είσοδος. Στην συνέχεια, το MapReduce συγκεντρώνει όλα τα ζεύγη και τα ομαδοποιεί ανά κλειδί. Η ομάδα αυτή θα δοθεί ως είσοδος στην συνάρτηση reduce.

$Reduce: : (key_2, list(value_2)) \rightarrow list(value_3)$

Κάθε κόμβος του cluster θα τρέξει την reduce στην ομάδα που του έχει ανατεθεί. Στο τέλος τα αποτελέσματα των reducers θα μαζευτούν και θα οδηγήσουν στο τελικό αποτέλεσμα. Λόγω του παραλληλισμού, τα δεδομένα θα εμφανιστούν με τυχαία σειρά, εκτος εάν ο προγραμματιστής επιλέξει να συμβεί η συσσώρευση με διαφορετικό τρόπο.

5.2.2 Hadoop

Το Hadoop αποτελεί την πιο διαδεδομένη υλοποίηση του MapReduce και είναι λογισμικό ανοιχτού κώδικα. Είναι υλοποιημένο σε Java και υποστηρίζει διαφορετικές γλώσσες, με πιο δημοφιλείς τις Java, Python, Scala. Μπορεί να διαχειριστεί μέχρι και petabytes δεδομένων ενώ μπορεί να λειτουργήσει αποδοτικά σε περισσότερους από 10.000 κόμβους.

Το εσωτερικό σύστημα αρχείων που χρησιμοποιεί το Hadoop ονομάζεται HDFS και είναι βασισμένο στην αρχιτεκτονική master/slave. Για ένα HDFS απαραίτητα είναι τα ακόλουθα :

- **Datanodes**. Τα δεδομένα είναι αποθηκευμένα στα Datanodes και είναι χωρισμένα σε blocks. Στην γενική περίπτωση, ένα Datanode αντιστοιχεί σε ένα μηχάνημα. Για λόγους ασφάλειας, υπάρχουν αντίγραφα όλων των δεδομένων εντός αυτών.

- **Namenode**. Απαραίτητη είναι η ύπαρξη ενώς Namenode ο οποίος έχει τον ρόλο του master. Αυτός διαχειρίζεται όλο το filesystem διαθέτοντας ευρετήριο για την τοποθεσία κάθε μπλοκ.

- **JobTracker**. Ο JobTracker είναι υπεύθυνος για την οργάνωση της εκτέλεσης των MapReduce εργασιών. Ενδεικτικά, προτεραιότητα του είναι να αναθέσει τις δουλειές σε κόμβους που διαθέτουν ή βρίσκονται κοντά σε κόμβους που διαθέτουν τα δεδομένα υπό επεξεργασία.

- **TaskTracker**. Οι TaskTrackers είναι αυτοί που εκτελούν τις εργασίες που ανατίθενται από τον JobTracker. Είναι υποχρέωση τους να αποστέλλουν συνεχόμενα παλμό ότι η λειτουργία τους συνεχίζει κανονικά, καθώς και να στέλνουν αναφορά επιτυχίας ή αποτυχίας όταν ολοκληρωθεί η δουλειά που τους ανατέθηκε.

Τα πλεονεκτήματα του HDFS συνοπτικά είναι τα ακόλουθα :

- Υψηλή ανοχή σε σφάλματα υλικού. Έτσι με το που προκύψει, εντοπίζονται οι κόμβοι που δημιούργησαν την αστοχία και γίνεται ανάκαμψη του συστήματος.

- Το HDFS είναι φτιαγμένο με σκοπό την επίτευξη υψηλού throughput.

- Εύκολη διαχείριση τεράστιων αρχείων.

- Εύκολη μεταφερσιμότητα σε διαφορετικές πλατφόρμες.

5.3 Προ-επεξεργασία Δεδομένων

5.3.1 Αρχική Μορφή

Η αρχική μορφή των δεδομένων ήταν αρκετά περίπλοκη και επομένως η προ-επεξεργασία ήταν πολύ σημαντικό κομμάτι. Παρακάτω παρουσιάζεται η μορφή των δεδομένων. Αρχικά πρέπει να αναφερθεί ότι στα δεδομένα υπάρχουν 3ων ειδών διαφορετικές εγγραφές, των οποίων η μορφή παρουσιάζεται παρακάτω :

- **Εγγραφή Τύπου M** : SessionID TypeOfRecord Day USERID
- **Εγγραφή Τύπου Q** : SessionID TimePassed TypeOfRecord SERPID QueryID ListOfTerms ListOfURLsAndDomains
- **Εγγραφή Τύπου C** : SessionID TimePassed TypeOfRecord SERPID URLID

SessionID είναι ένας μοναδικός αριθμός για να ταυτοποιείται κάθε συνεδρία αναζητήσεων.

Day είναι η ημέρα που πραγματοποιήθηκε η συγκεκριμένη ενέργεια (τα δεδομένα αφορούν μία περίοδο 30 ημερών).

TypeOfRecord είναι ο τύπος της εγγραφής. Είναι είτε αναζήτηση κάποιας φράσης (Q), είτε κλικ (C), είτε εγγραφή που περιέχει γενικά στοιχεία για μία συνεδρία όπως η μέρα και το ID του χρήστη (M).

Το UserID είναι ο μοναδικός αριθμός που ταυτοποιεί τον χρήστη.

TimePassed είναι ο χρόνος ανάμεσα στην σύνδεση του χρήστη και στην επιλογή της εκάστοτε σελίδας. Ο χρόνος μετριέται σε μονάδες που δεν γίνεται γνωστό πόσα ms περιέχουν.

QueryID είναι το μοναδικό αναγνωριστικό που κωδικοποιεί την συγκεκριμένη αναζήτηση που έκανε ο χρήστης.

ListOfTerms είναι μία λίστα από τους όρους- λέξεις που αναζήτησε ο χρήστης. Κάθε λέξη παρουσιάζεται σε μορφή ID.

SERPID είναι το ID της σελίδας που εμφανίστηκε ως αποτέλεσμα στον χρήστη.

TermId είναι το ID της φράσης που αναζήτησε ο χρήστης.

Το URLID είναι το μοναδικό ID του εκάστοτε URL.

Το ListOfURLsAndDomains είναι μία λίστα από ζευγάρια URLID και DomainID χωρισμένων με υποδιαστολή και αφορούν τα αποτελέσματα στην αναζήτηση του χρήστη. Η σειρά με την οποία δίνονται είναι και η σειρά με την οποία εμφανίστηκαν τα αποτελέσματα, από πάνω προς τα κάτω. Ανάμεσα σε κάθε πλειάδα (tuple), υπάρχει TAB.

Παράδειγμα

744899 M 23 123123123

744899 0 X 0 192902 4857, 3847, 2939 632428,2384 309585,28374 319567,38724 6547,28744 20264,2332 3094446,34535 90,21 841,231 8344,2342 119571,45767

744899 1403 " 0 632428

Έτσι, σύμφωνα με τα παραπάνω, ο χρήστης 123123123, ξεκίνησε ένα session 744899, το οποίο συνέβη την μέρα νούμερο 23. Σε αυτή την αναζήτηση αναζήτησε τις 3 λέξεις 4857,3847,2939 με ID αναζήτησης 192902. Τελικά του δόθηκε μία σειρά από URLs. Μετά από 1403 μονάδες χρόνου, ο χρήστης επέλεξε το URL 632428, το οποίο είχε domain 2384 και βρισκόταν στην πρώτη θέση των αποτελεσμάτων (όπως φαίνεται από την εγγραφή Q).

5.3.2 Προ-επεξεργασία

Όπως έγινε κατανοητό τα δεδομένα βρίσκονται σε δύσκαμπτη μορφή και απαιτείται καλύτερη οργάνωση για να μπορούν να βγουν συμπεράσματα. Το πρώτο βήμα αφορά την πληροφορία η οποία δεν μας ήταν χρήσιμη και η οποία αποφασίστηκε να μην χρησιμοποιηθεί. Πιο συγκεκριμένα :

- Τα δεδομένα σχετικά με τον χρήστη αποφασίστηκε ότι δεν μπορούν να φανούν χρήσιμα, αφού μιλάμε για πολύ λίγη πληροφορία για τον κάθε χρήστη ξεχωριστά. Ενδεικτικά, το σύστημα συμπεριλαμβάνει 5 εκατομμύρια διαφορετικούς χρήστες σε 65 εκατομμύρια διαφορετικές αναζητήσεις. Το ίδιο συνέβη με την ημέρα καθώς δεν υπήρχε κάποια χρηστική αξία σε αυτήν. Επομένως όλες οι εγγραφές τύπου M δεν χρησιμοποιήθηκαν.

- Επειδή στις εγγραφές δεν υπάρχει ένα στοιχείο το οποίο να μπορεί να χρησιμοποιηθεί ως κλειδί - μοναδικό αναγνωριστικό, δημιουργήθηκε ένα. Αυτό έγινε με την συγχώνευση 2 στοιχείων, του SessionID και του SERPID. Η συγχώνευση συνέβη προσθέτοντας μία παύλα ανάμεσα τους ώστε κάθε SessionID-SERPID να είναι μοναδικό και να λειτουργεί σαν μοναδικό κλειδί στο μέλλον.

- Η πληροφορία σχετική με τον χρόνο επίσης δεν μπορούσε να φανεί χρήσιμη στην ανάλυση μας και γι΄ αυτό δεν χρησιμοποιήθηκε.

- Δυστυχώς η κωδικοποίηση των λέξεων σε αριθμητική μορφή δεν βοήθησε την ανάλυση με χρήση τεχνικών Μηχανικής Μάθησης σε αυτές και γι΄ αυτό δεν χρησιμοποιήθηκαν. Ενδεχομένως να μπορούσε να γίνει ανάλυση σε περίπτωση που δεν δίνονταν οι λέξεις αλλά οι κατηγορίες αναζητήσεων. Αλλά στην μορφή την οποία δόθηκαν δεν μπορούσαν να χρησιμοποιηθούν, ενδεικτικά είχαμε 5 εκατομμύρια διαφορετικές λέξεις που αναζητήθηκαν. Για αντίστοιχους λόγους δεν χρησιμοποιήθηκε ούτε τα Query IDs αφού ήταν 21 εκατομμύρια διαφορετικά, σε 65 εκατομμύρια διαφορετικές αναζητήσεις, αναλογία η οποία δεν τα καθιστά χρηστικά.

- Τέλος, ο τύπος εγγραφής δεν χρειάζεται στα δεδομένα αφού μπορούμε να καταλάβουμε τον τύπο εγγραφής από τον αριθμό των στοιχείων της εγγραφής.

Έτσι, μετά τα παραπάνω, το παράδειγμα που δόθηκε αρχικά διαμορφώνεται ως εξής :

744899-0 632428,2384 309585,28374 319567,38724 6547,28744 20264,2332 3094446,3453590,21 841,231 8344,2342 119571,45767

744899-0 632428

Το επόμενο βήμα αφορά την επεξεργασία των URLs και των domains. Αποφασίστηκε ότι η επεξεργασία που θα συμβεί θα αφορά τα domains. Η επιλογή αυτή γίνεται για έναν αριθμό από λόγους. Πρώτον, μπορεί να υποθέσει κανείς ότι κάθε σελίδα του ίδιου ιστότοπου έχει την ίδια περίπου ποιότητα. Επομένως αφού δεν μπορεί να παρατηρηθεί ανάμεσα στις διαφορετικές σελίδες μεγάλη διακύμανση, τότε μπορεί να γίνει η γενίκευση ότι κάθε σελίδα του ίδιου ιστότοπου έχει την ίδια ποιότητα. Δεύτερον, η ανάλυση των domains μας δίνει πιο μεγάλη συνοχή στην έρευνα, καθώς το μεγαλύτερο μέρος των URLs θα συναντάται μία μόνο φορά και επομένως δεν θα υπάρχουν επαρκή στατιστικά.

Το πρόβλημα που προκύπτει από την επιθυμία να γίνει η ανάλυση με domains είναι πως οι εγγραφές τύπου C έχουν το επιλεγμένο URL και όχι το domain. Επομένως με κάποιον τρόπο έπρεπε να γίνει η μεταφορά από URL σε domain. Με την βοήθεια του MapReduce, το πρόβλημα αυτό λύθηκε γρήγορα και αποδοτικά χρησιμοποιώντας ως κλειδί συγχωνευμένα το κάθε URL ξεχωριστά μαζί με τα μοναδικά κλειδιά. Αφού πλέον είχαμε κρατήσει το domain κάθε κλικ, οι πληροφορίες σχετικά με τα URLs δεν κρατήθηκαν. Επομένως πλέον οι εγγραφές ήταν της ακόλουθης μορφής :

744899-0 2384 28374 38724 28744 2332 34535 21 231 2342 45767

744899-0 2384

Στο επόμενο βήμα έγιναν 2 χρήσιμες αλλαγές. Οι εγγραφές του κάθε session συγκεντρώθηκαν στην ίδια εγγραφή (τα κλικ προστέθηκαν στο ακριβώς μετά τα domains των αποτελεσμάτων) και προστέθηκε ως τελευταίο στοιχείο της εγγραφής ένας αριθμός ο οποίο έδειχνε μέχρι ποιό σημείο έφτασε να βλέπει ο χρήστης. Αφού τα μοντέλα που σκοπεύουμε να εξετάσουμε είναι σειριακά (δηλαδή ο χρήστης βλέπει τα URLs στην σειρά), υποθέσαμε ότι ο χρήστης σταμάτησε να εξετάζει URLs στο τελευταίο URL που έκανε κλικ. Πρόκειται για μία υπόθεση η οποία παρ΄ ότι δεν αντιπροσωπεύει ολοκληρωτικά την πραγματικότητα, ήταν απαραίτητο να γίνει καθώς δεν υπήρχε διαφορετικός τρόπος υπολογισμού του συγκεκριμένου στατιστικού.

Επομένως το παράδειγμα που δόθηκε παραπάνω πλέον γίνεται :

744899-0 2384 28374 38724 28744 2332 34535 21 231 2342 45767 2384 1

Επομένως ο χρήστης έφτασε μέχρι το URL 2384, έκανε κλικ σε αυτό και σταμάτησε. Εάν για παράδειγμα είχε κάνει 2 κλικ, στο πρώτο και στο 5ο URL, η εγγραφή θα ήταν :

744899-0 2384 28374 38724 28744 2332 34535 21 231 2342 45767 2384 2332 5

Πλέον η μορφή είναι ικανοποιητική για να εξαχθούν αποτελέσματα σχετικά με τα domains. Πριν από αυτό θα πρέπει να δοθεί ένας μικρός ορισμός λεξιλογικός για να γίνει με μεγαλύτερη ακρίβεια η περιγραφή της συνέχειας.

Ορισμός 5.1. ***Εμφάνιση***

Ως εμφάνιση ορίζουμε την περίπτωση που μία σελίδα συμπεριλήφθηκε στην σελίδα ανεξάρτητα απ΄ το εάν επιλέχθηκε, εάν ο χρήστης την εξέτασε αλλά δεν την επέλεξε ή ο χρήστης δεν έφθασε καν ποτέ να δει την σελίδα αυτή.

Ορισμός 5.2. ***Προβολή***

Ως προβολή ορίζουμε την περίπτωση που μία σελίδα συμπεριλήφθηκε στην σελίδα και ο χρήστης την εξέτασε για το εάν θα την ανοίξει ή όχι.

Ορισμός 5.3. ***Συνέχεια***

Ως συνέχεια ορίζουμε την περίπτωση που μία σελίδα προβλήθηκε και ο χρήστης συνέχισε να προβάλει επόμενες σελίδες.

Τώρα συνεχίζουμε με την επεξεργασία των δεδομένων. Για κάθε domain κρατήσαμε τα ακόλουθα στατιστικά :

- Αριθμός εμφανίσεων.
- Αριθμός προβολών.
- Αριθμός κλικ.
- Αριθμός φορών συνέχειας.

Το τελευταίο στατιστικό στοιχείο παρ΄ ότι μοιάζει ασυνήθιστο, αφορά την πιθανότητα συνέχειας που ορίστηκε στο 3ο κεφάλαιο και είναι πολύ σημαντικό στατιστικό στοιχείο για το μοντέλο Αλληλουχίας. Πριν συνεχίσουμε πρέπει να γίνουν **δύο πολύ σημαντικές παρατηρήσεις** :

- Για τις διαφημίσεις που δεν υπάρχουν αρκετά δεδομένα (ορίστηκαν ως κατώφλι οι 20 προβολές), τότε θεωρήθηκε πως η πιθανότητα συνέχειας είναι 1 ενώ οι υπόλοιπες πιθανότητες 0.
- Σε αυτό το σημείο έγινε διαχωρισμός των δεδομένων. Για τον υπολογισμό των στατιστικών σχετικά με τα domains, χρησιμοποιήθηκαν τα μισά δεδομένα για τα οποία ήταν η τελευταία φορά που χρησιμοποιήθηκαν. Η συνέχεια γίνεται με το άλλο 25% των δεδομένων ενώ ένα τελευταίο 25% έμεινε αχρησιμοποίητο σε περίπτωση που προκύψουν καινούριες ανάγκες στο πρόβλημά μας.

Πλέον με τους αριθμούς που διαθέτουμε έχουμε ότι στατιστικό στοιχείο έχουμε ανάγκη για τα domains. Μάλιστα υπολογίζονται οι λόγοι :

κλικ/προβολές , εμφανίσεις/προβολές , συνέχεια/προβολές

Στα επόμενα στάδια συμβαίνουν 2 αλλαγές, αρχικά σε κάθε εγγραφή προστίθεται εάν ο χρήστης έκανε κλικ σε δυαδική μορφή και αφαιρούνται τα domains των κλικ από το τέλος. Έτσι, το γνωστό μας παράδειγμα γίνεται :

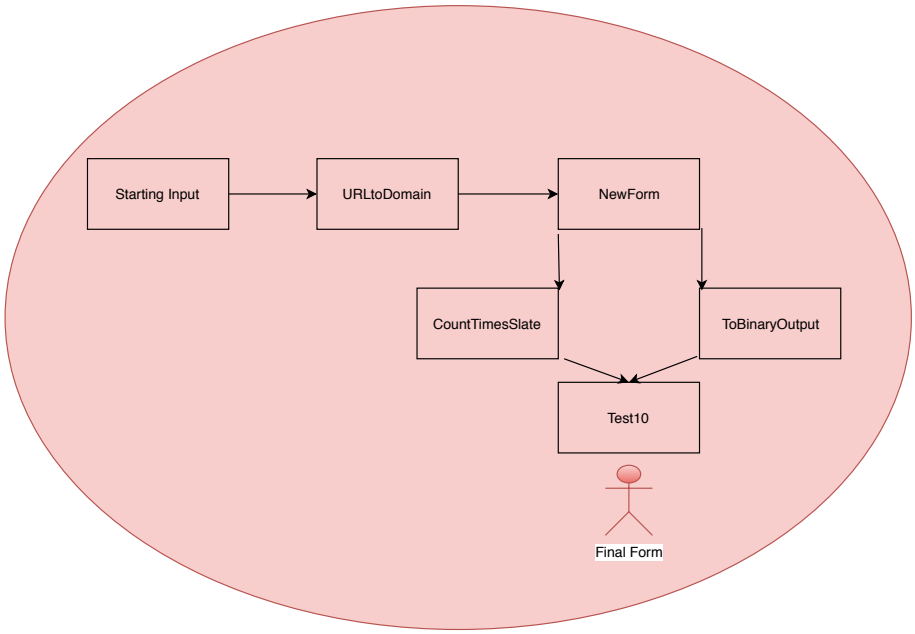
744899-0 1 0 0 0 0 0 0 0 0 0 2384 28374 38724 28744 2332 34535 21 231 2342 45767

Το οποίο σημαίνει πως ο χρήστης έκανε κλικ μόνο στο 1ο domain.

Πλέον βρισκόμαστε σε θέση να αφαιρέσουμε την πληροφορία των domains και να τα αντικαταστήσουμε με τα 3 στατιστικά του κάθε domain. Έτσι, η μορφή μας πλέον είναι :

744899-0 1 0 0 0 0 0 0 0 0 0 0.42,0.61,0.81 0.23,0.4,0.9 ...

Με τα παραπάνω βήματα η προ-επεξεργασία τελείωσε και σειρά είχε η ανάλυση. Για να συνοψιστεί η προ-επεξεργασία σε ένα σχήμα, όλα όσα περιγράφηκαν παραπάνω βρίσκονται στο ακόλουθο διάγραμμα. Να σημειωθεί πως τα ονόματα στα κουτιά, είναι ακριβώς τα ονόματα του εκάστοτε κώδικα που μπορείτε να βρείτε στο gitHub repository [30].



Σχήμα 5.2: Η διαδικασία της προ-επεξεργασίας δεδομένων

Συνοψίζοντας, από την αρχική μορφή καταλήγουμε σε μία λίστα από κλικ με τα URLs να έχουν μετατραπεί σε domains, το αρχείο URLtoDomain. Συνδυάζοντας την αρχική μορφή με το προηγούμενο, στο NewForm καταφέρνουμε να διώξουμε την πληροφορία που δεν χρειαζόμαστε και να συνδυάσουμε όλη την πληροφορία που διαθέτουμε μαζί σε μία εγγραφή, προσθέτοντας τον αριθμό που δείχνει το σημείο στο οποίο σταμάτησε να βλέπει ο χρήστης. Στο CountTimesSlate δημιουργούμε μία λίστα με κάθε domain και τα στατιστικά του ενώ στο ToBinaryOutput προσθέτουμε την δυαδική κωδικοποίηση. Τέλος καταφέρνουμε να διώξουμε ολοκληρωτικά τα domains, κρατώντας μόνο τα στατιστικά σχετικά με αυτά και με αυτόν τον τρόπο φτάνουμε στην τελική μας μορφή.

Ειδικές Περιπτώσεις

Θα αναφερθούν παρακάτω κάποιες περιπτώσεις που απαιτούσαν ειδική διαχείριση.

- **Δύο κλικ στο ίδιο URL**. Ορισμένες φορές οι εγγραφές περιείχαν δύο κλικ στο ίδιο URL, γεγονός το οποίο εάν δεν είχαν ειδική μεταχείριση, θα μέτραγε ως 2 κλικ. Στην πραγματικότητα όμως αυτό έπρεπε να μετρά ως 1 κλικ και γι΄ αυτό χρειάστηκε να γίνει διαγραφή των 2ων κλικ.

- **Δύο ίδιες αναζητήσεις**. Πολλές φορές οι χρήστες για να βρουν την αναζήτηση που έκαναν προηγουμένως, επέστρεφαν στην σελίδα αποτελεσμάτων ξανακάνοντας την ίδια αναζήτηση. Το γεγονός αυτό καταχωρούνταν στα δεδομένα ως 2 διαφορετικές

εγγραφές. Στο σημείο αυτό επιλέχθηκε οι 2 αναζητήσεις να συγχωνευτούν σε μία καθώς ο χρήστης ενδιαφερόταν να επιλέξει κάποιες σελίδες ακόμα από την προηγούμενή του αναζήτηση.

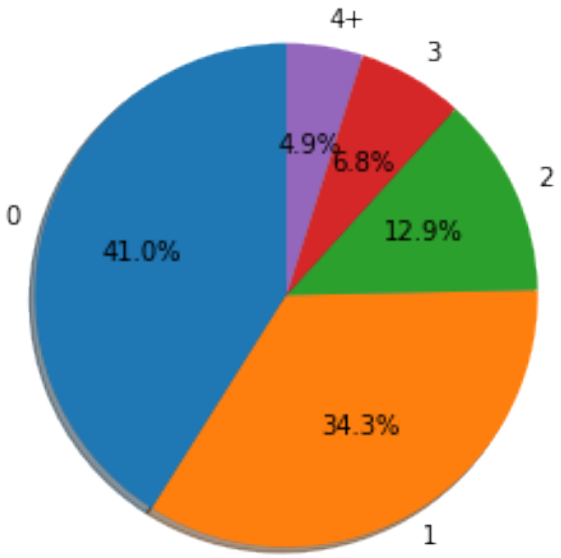
5.4 Ανάλυση Δεδομένων

Στην μορφή που έχουμε πλέον τα δεδομένα μας, μπορούμε να συνεχίσουμε με την ανάλυση των δεδομένων μας. Αρχικά να δοθούν κάποια στατιστικά για τα δεδομένα τα οποία υπολογίστηκαν για να υπάρχει καλύτερη εικόνα :

Γενικά Στατιστικά

- 5 εκατομμύρια διαφορετικές λέξεις
- 65 εκατομμύρια αναζητήσεις
- 21 εκατομμύρια διαφορετικές αναζητήσεις
- 5 εκατομμύρια διαφορετικοί χρήστες
- 700 χιλιάδες διαφορετικά URLs
- 64 εκατομμύρια κλικ

Αριθμός κλικ ανά σελίδα



Σχήμα 5.3: Αριθμός κλικ ανά σελίδα

Όπως φαίνεται παραπάνω, το 41% των χρηστών δεν επιλέγει καμία σελίδα, το 34.3% επιλέγει μία ενώ το 4.5% επιλέγει 4 και πάνω. Επίσης υπολογίστηκε ότι **ο μέσος όρος του αριθμού των κλικ βρίσκεται πολύ κοντά στο 1 κλικ ανά σελίδα**. Πιο

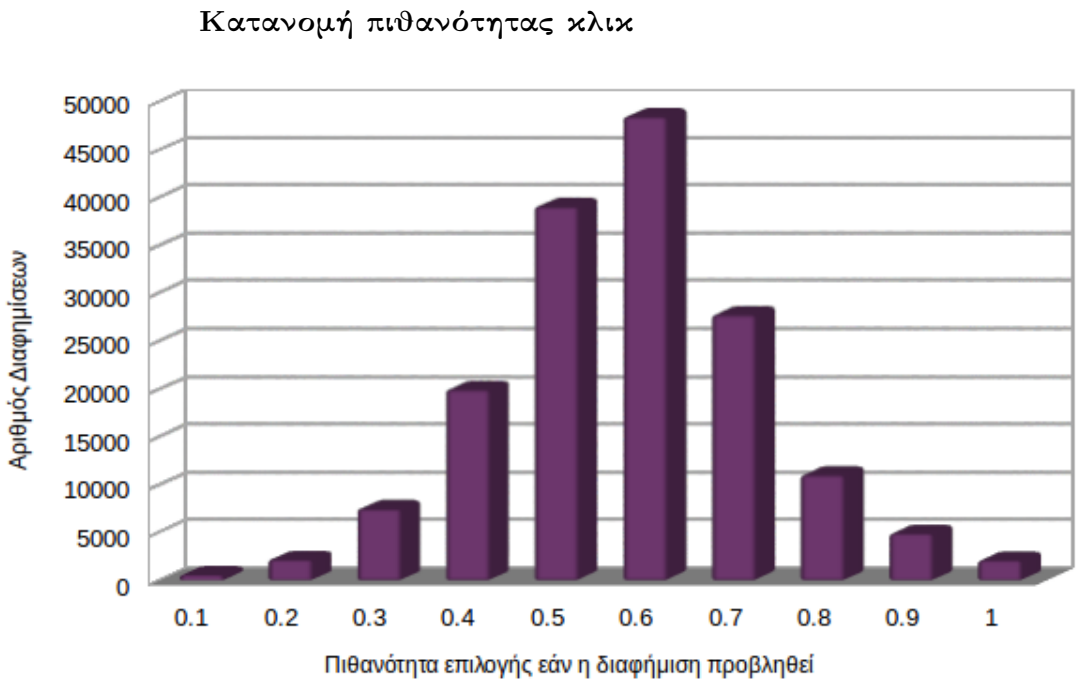
συγκεκριμένα το 89.5% δεν επιλέγονται. Επειδή κάθε σελίδα έχει 10 URL, έχουμε κατά μέσο όρο περίπου 1 κλικ ανά σελίδα.

Πιθανότητα κλικ ανά θέση

Σχήμα 5.4: Πιθανότητα κλικ ανά θέση

Πιθανότητα προβολής ανά θέση

Σχήμα 5.5: Πιθανότητα προβολής ανά θέση



Σχήμα 5.6: Κατανομή πιθανότητας κλικ

Σχήμα 5.7: Κατανομή πιθανότητας συνέχειας

Από τα παραπάνω διαγράμματα, ιδιαίτερη έμφαση πρέπει να δοθεί στο 2ο, καθώς οι αριθμοί που υπολογίστηκαν είναι οι πολλαπλασιαστές σχετικοί με την θέση που ορίζονται στο διαχωρίσιμο στο 3ο κεφάλαιο. Η 3η και 4η γραφική δείχνουν την κατανομή που ακολουθούν

οι πιθανότητες συνέχειας και κλικ των διαφημίσεων. Στην γραφική δεν έχει συμπεριληφθεί η πιθανότητα 0 γιατί λόγω των περιορισμένων στατιστικών, είναι πολύ μεγάλος ο αριθμός των διαφημίσεων που έχουν πιθανότητα 0 και επομένως εάν είχε συμπεριληφθεί στο ιστόγραμμα, δεν θα φαινόταν καλά η κλιμάκωση των μεγεθών. Με τα παραπάνω αποκτήθηκε μία πολύ καλή εικόνα των δεδομένων και υπολογίστηκαν ότι μεγέθη χρειάστηκαν για την υλοποίηση των μοντέλων της βιβλιογραφίας.

5.5 Αξιολόγηση Μοντέλων

5.5.1 Σύγκριση Αλληλουχίας-Διαχωρίσιμου Μοντέλων

Το πρώτο βήμα ήταν η σύγκριση των δύο μοντέλων της βιβλιογραφίας. Δημιουργήθηκαν 2 κώδικες, οι Separable.java - Cascade.java, οι οποίοι υπολογίζουν τις πιθανότητες κλικ του χρήστη για κάθε σελίδα. Στην συνέχεια χρησιμοποιήθηκαν οι μετρικές για regression που περιγράφτηκαν στο κεφάλαιο 4, ώστε να βγάλουμε το συμπέρασμα ποιό από τα μοντέλα είναι καλύτερο. Τα αποτελέσματα αυτά φαίνονται στον παρακάτω πίνακα :

Αλληλουχίας vs Διαχωρίσιμο		
Μετρική	Αλληλουχίας	Διαχωρίσιμο
RMSE	0.29205726266166215	0.2802298259204328
Norm 1	0.17464435306935447	0.14583345793964328
R^2	0.6148247785936977	0.8793513251549624

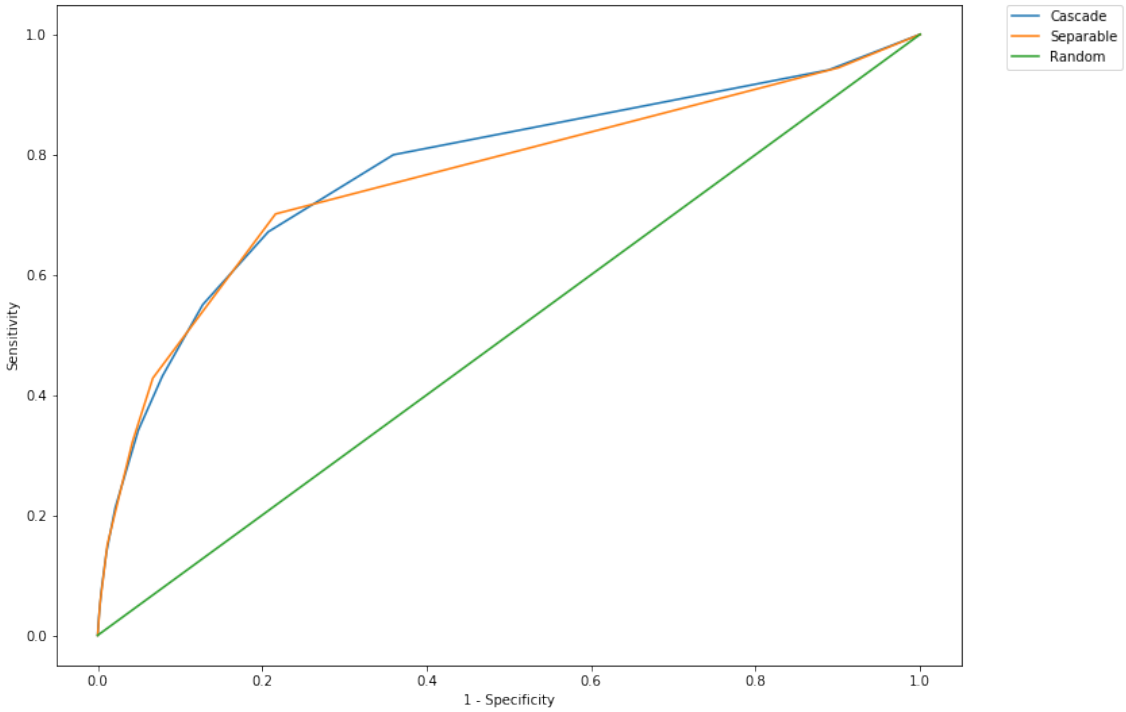
Όπως παρατηρούμε, τα 2 μοντέλα βρίσκονται πολύ κοντά, με το διαχωρίσιμο μοντέλο να είναι λίγο καλύτερο, γεγονός το οποίο με πρώτη ματιά μοιάζει αξιοπερίεργο καθώς θεωρήθηκε ως το πιο απλό μοντέλο ενώ το Αλληλουχίας θεωρήθηκε πιο πλούσιο. Φυσικά, λόγω των labels 0 και 1 που έχουμε, έχουμε ένα πρόβλημα classification. Έτσι, συνεχίζουμε με την μετατροπή του προβλέπτη μας σε πρόβλημα classification. Για περισσότερη πληροφορία έγινε η ακόλουθη δοκιμή, δημιουργήθηκε ένας confusion matrix. Ο confusion matrix είναι μία τεχνική που χρησιμοποιείται σε classification προβλήματα και δείχνει ποιές εγγραφές δεδομένων προβλέφθηκαν σωστά ως σωστές **True Positive-TP**, ποιές σωστά ως λάθος **True Negative-TN**,ποίες λανθασμένα ως σωστές **False Positive-FP** και ποίες λανθασμένα ως λάθος **False Negative-FN**. Ως κατώφλι θέσαμε το 0.5 και το αποτέλεσμα που πήραμε βρίσκεται παρακάτω :

Αλληλουχίας vs Διαχωρίσιμο, Confusion Matrixes		
Κατηγορία	Αλληλουχίας	Διαχωρίσιμο
TP	2215792	345074
FP	4280419	6151137
FN	2741864	157948
TN	52815705	55399621

Ο παραπάνω πίνακας μας δίνει περισσότερα στοιχεία. Η παρατήρηση που κάνουμε είναι ότι το μοντέλο Αλληλουχίας είναι πολύ καλύτερο στο να προβλέπει εύστοχα τα κλικ του χρήστη. Αντίθετα το Διαχωρίσιμο προβλέπει καλύτερα τα μη κλικ. Η ιδιότητα αυτή του μοντέλο Αλληλουχίας μας δίνει ήδη ένα σοβαρό πλεονέκτημα σε σχέση με το Διαχωρίσιμο το οποίο μας ωθεί στην περαιτέρω αναζήτηση των ιδιοτήτων των 2 μοντέλων. Πλέον, από το confusion matrix μπορούμε να υπολογίσουμε το accuracy του μοντέλου το οποίο φαίνεται στον παρακάτω πίνακα :

Αλληλουχίας vs Διαχωρίσιμο, Confusion Matrixes			
	Αλληλουχίας	Διαχωρίσιμο	Προβλέπτης πάντα 0(μη κλικ)
Accuracy	90.02%	90.05%	89.53%

Τέλος για περισσότερη πληροφορία χαράσσουμε τις γραφικές παραστάσεις ROC.



Σχήμα 5.8: Αλληλουχίας vs Διαχωρίσιμο, ROC curve

Στην παραπάνω γραφική φαίνεται φανερά το πλεονέκτημα του μοντέλου Αλληλουχίας. Το πλεονέκτημα του είναι πως **περιγράφει με μεγαλύτερη ακρίβεια τα κλικ του χρήστη** σε σχέση με το Διαχωρίσιμο. Παρακάτω υπολογίζεται το εμβαδόν κάτω από την καμπύλη (AUC).

AUC Αλληλουχίας vs Διαχωρίσιμο		
	Αλληλουχίας	Διαχωρίσιμο
AUC	77.48%	76.14%

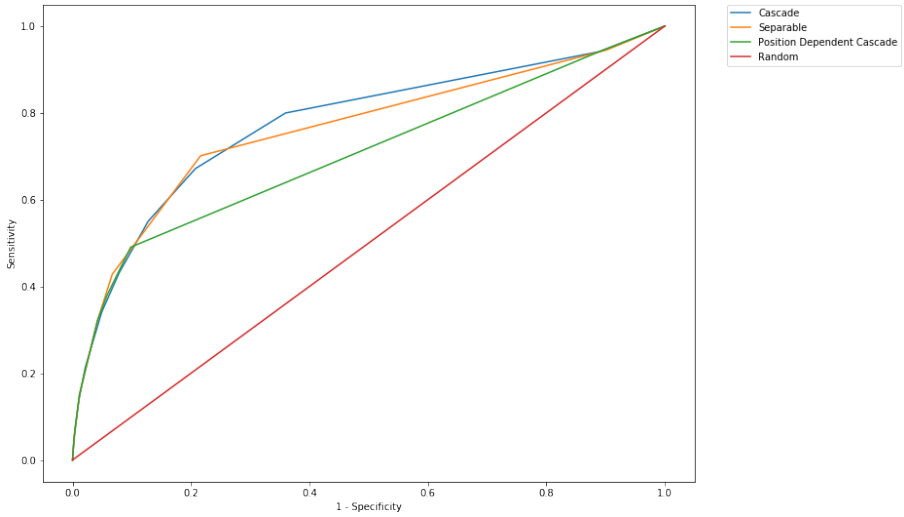
Συμπέρασμα

Πρόκειται για δύο μοντέλα τα οποία βρίσκονται πολύ κοντά στην πρόβλεψη της συμπεριφοράς του χρήστη. Παρ΄ ότι στις περισσότερες μετρικές το διαχωρίσιμο αποδίδει καλύτερα, το μοντέλο Αλληλουχίας κάνει πιο ακριβή πρόβλεψη των κλικ του χρήστη. Όμως, λόγω της πιο ακριβούς πρόβλεψης των τιμών της συμπεριφοράς του χρήστη όταν δεν κάνει κλικ, οι οποίες αποτελούν το 90% των περιπτώσεων, το διαχωρίσιμο βρίσκεται ελάχιστα πιο πάνω στην ακρίβεια.

5.5.2 Εναλλακτικά Μοντέλα

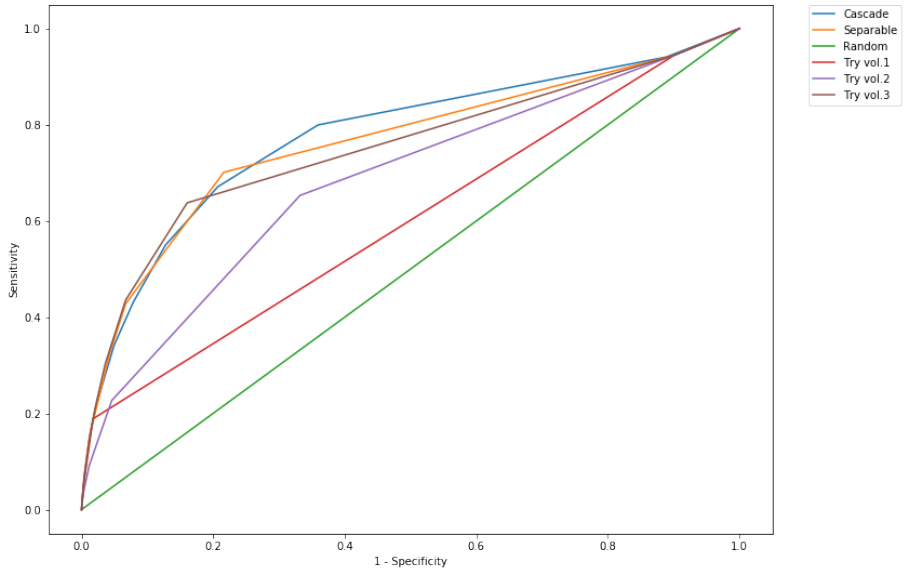
Στην συνέχεια δοκιμάστηκαν διάφορα μοντέλα με στόχο την εύρεση ενός μοντέλου που ανταγωνίζεται ή και ξεπερνά τα 2 μοντέλα της βιβλιογραφίας. Παρακάτω θα παρουσιαστούν οι δοκιμές που έγιναν και δεν απέδωσαν ικανοποιητικά. Στο τέλος δίνεται μία ονομασία ώστε να μπορεί να γίνει σύνδεση του μοντέλου με την αντίστοιχη γραφική παράσταση. Σε κάθε γραφική παράσταση για να υπάρχει εικόνα παρουσιάζονται οι 2 ευθείες του διαχωρίσιμο και του μοντέλου Αλληλουχίας, καθώς και η ευθεία της τυχαίας επιλογής.

- Η πρώτη δοκιμή αφορούσε μία προσπάθεια να συνδυαστούν και τα 2 μοντέλα σε ένα.[;] Έτσι η πιθανότητα να κάνει κλικ ο χρήστης υπολογίζεται από τον τύπο : $q_{a_i} \cdot \lambda_i \cdot \prod_{j=1}^{i-1} c_{a_j}$. Το συγκεκριμένο μοντέλο δεν απέδωσε καλά καθώς οι πολλαπλασιαστές πρέπει να υπολογιστούν με πειραματική αξιολόγηση από την αρχή. (Position Dependent Cascade)



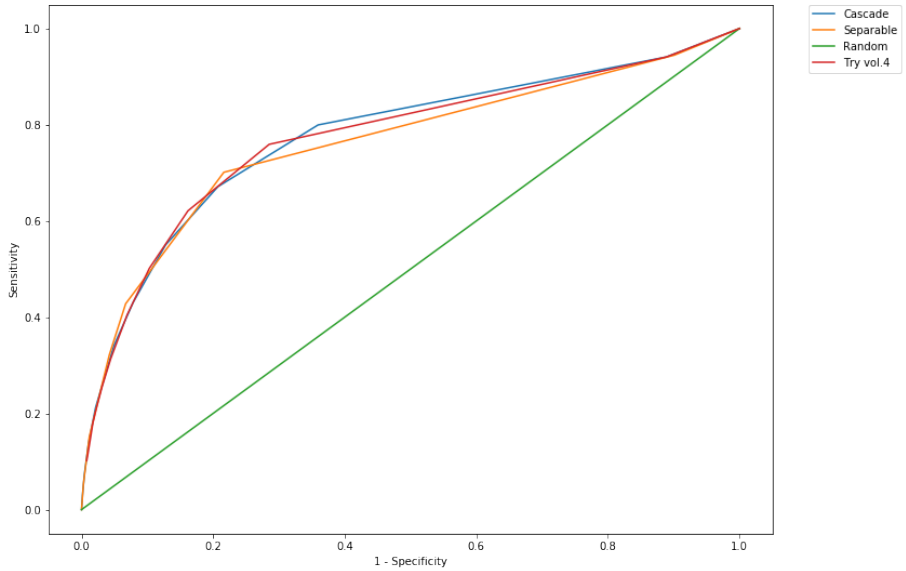
Σχήμα 5.9: Position Dependent Cascade

- Στην συνέχεια έγινε προσπάθεια να υπολογιστεί με διαφορετικό τρόπο το q_{a_i} των μοντέλων. Αντί για τον λόγο κλικ ανά προβολές δοκιμάστηκε ο λόγος κλικ ανά εμφανίσεις(Try vol.1, Try vol.2) και ο λόγος κλικ ανά εμφανίσεις κανονικοποιημένο στο μέγεθος του λόγου κλικ ανά προβολές (Try vol.3). Τα παραπάνω δοκιμάστηκαν όλα με βάση το διαχωρίσιμο μοντέλο, με το Try vol.3 να αποδίδει αρκετά καλά, καταφέρνοντας για μικρά κατώφλια να ξεπεράσει τα 2 μοντέλα της βιβλιογραφίας και τελικά να έχει accuracy 89.99%.



Σχήμα 5.10: Παραλλαγές Διαχωρίσιμο Μοντέλου

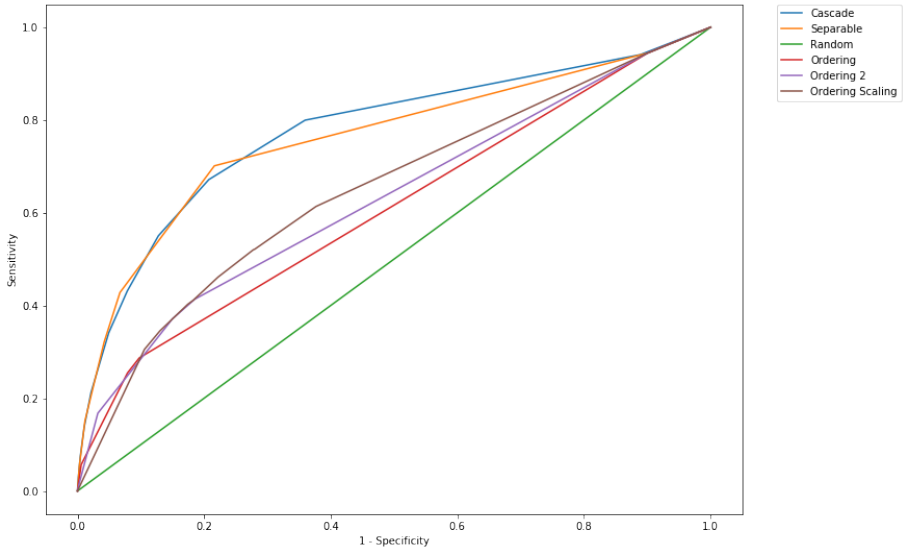
- Η αντίστοιχη προσπάθεια με το Try vol.3 έγινε για το μοντέλο Αλληλουχίας, η οποία απέδωσε και πάλι πολύ κοντά στα μοντέλα της βιβλιογραφίας και ακρίβεια 89.88%.



Σχήμα 5.11: Παραλλαγή Μοντέλου Αλληλουχίας

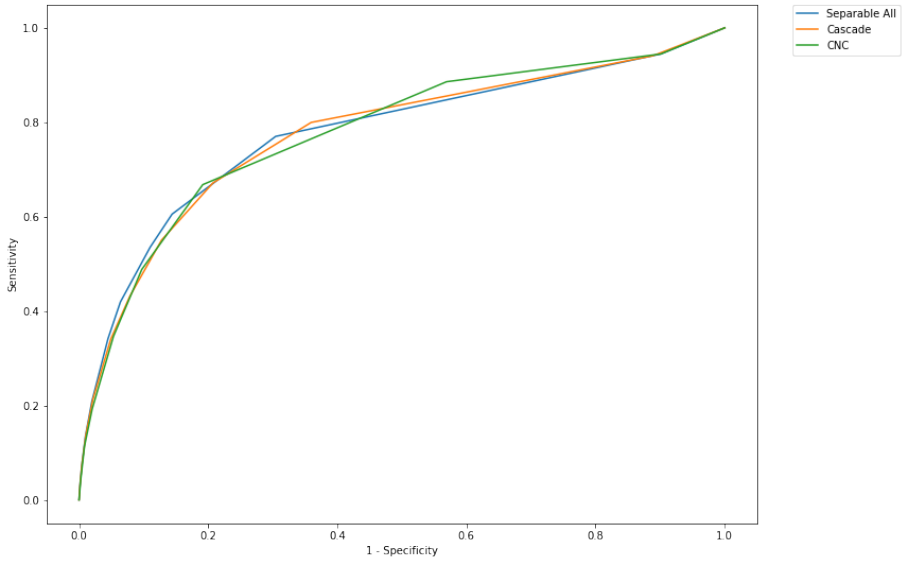
- Οι επόμενες προσπάθειες ήταν μοντέλα βασισμένα στην ιδέα ότι ο χρήστης δεν διαβάζει τα αποτελέσματα με σειρά θέσεων. Τα μοντέλα αυτά δεν απέδωσαν σε ανταγωνιστικό επίπεδο σε σχέση με τα δύο βασικά μοντέλα. Το CasDifo ήταν μία προσπάθεια να υπολογιστεί η πιθανότητα συνέχειας από την δημοφιλέστερη επιλογή προς την λιγότερη δημοφιλή, ενώ τα επόμενα δύο υπολόγισαν την θέση κάθε διαφήμισης ταξινομώντας τις με σειρά δημοφιλέστερου λινκ και τελικά υπολογίστηκαν μέσω του διαχωρίσιμου

μοντέλου με την νέα ταξινομημένη σειρά (Ordering και Ordering Scaling).



Σχήμα 5.12: Μοντέλα Ταξινόμησης

- Τέλος, παρουσιάζεται η πιο ελπιδοφόρα προσπάθεια. Σύμφωνα με το μοντέλο αυτό (CNC), την πιθανότητα συνέχειας του εκάστοτε λινκ δεν είναι σταθερό αλλά προκύπτει από την πιθανότητα ο χρήστης να μην έχει επιλέξει καμία από τις προηγούμενες διαφημίσεις. Επομένως η πιθανότητα επιλογής ενός λινκ είναι : $r_{a_i} = q_{a_i} \cdot No$, όπου $No = \prod_{j=1}^{i-1} 1 - r_{a_j}$



Σχήμα 5.13: CNC

Το μοντέλο αυτό έχει εξαιρετική ROC curve και μάλιστα ξεπερνά τα δύο γνωστά μας μοντέλα στο εμβαδόν κάτω από την καμπύλη. Επιπλέον υπολογίστηκε το accuracy το

οποίο παρουσιάζεται κι αυτό στον επόμενο πίνακα.

CNC			
	Αλληλουχίας	Διαχωρίσιμο	CNC
AUC	77.48%	76.14%	77.72%
Accuracy	90.02%	90.05%	89.96%

Τέλος για λόγους πληρότητας παρουσιάζονται τα AUC όλων των μοντέλων που παρουσιάστηκαν παραπάνω.

All Models AUC	
	AUC
Αλληλουχίας	77.48%
Διαχωρίσιμο	76.14%
CNC	77.72%
Position Dependent Cascade	29.51%
Try vol.1	59.87%
Try vol.2	68.26%
Try vol.3	75.01%
Try vol.4	77.14%
CasofDif	60.75%
Ordering	63.06%
Ordering Scaling	64.92%

5.6 Ποιοτική Ανάλυση

Τα 3 μοντέλα βρίσκονται πολύ κοντά στην ακρίβεια με την οποία προβλέπουν την συμπεριφορά του χρήστη. Παρ΄ ότι τα μοντέλα βρίσκονται πολύ κοντά, παρατηρούμε πως η ακρίβεια είναι αντιστρόφως ανάλογη της ακρίβειας με την οποία προβλέπονται οι διαφημίσεις οι οποίες

είναι επιλεγμένες (έχει γίνει κλικ) από τον χρήστη. Το παραπάνω είναι απολύτως λογικό, ειδικά αφού το 89.5% των δεδομένων είναι στην κλάση όχι κλικ. Επομένως με σειρά ακρίβειας έχουμε : διαχωρίσιμο, Αλληλουχίας και CNC, ενώ το AUC έχει ακριβώς με την αντίστροφη σειρά.

Για την κατανόηση των αποτελεσμάτων, πρέπει να γίνουν κατανοητά τα ακόλουθα σημεία :

- Το Διαχωρίσιμο μοντέλο σχεδόν ταυτίζεται με το μοντέλο που προβλέπει ότι όλα είναι στην κλάση όχι κλικ (0), για τις θέσεις 5-10. Αυτό συμβαίνει καθώς οι πιθανότητες λ_i είναι πολύ μικρές και επομένως το r_i δεν περνά ποτέ το κατώφλι 0.5.
- Όμως, το Μοντέλο Αλληλουχίας αποδίδει χειρότερα από το διαχωριστικό καθώς κάνει περισσότερα λάθη στην ανάλυση των θέσεων 5-10.

Για την περαιτέρω κατανόηση των αποτελεσμάτων, γίνεται ανάλυση της ακρίβειας των μοντέλων ανά θέση. Πρώτα παρουσιάζονται τα αποτελέσματα και στην συνέχεια θα εξαχθούν τα συμπεράσματα.

Ανάλυση ανά θέση				
Θέσεις	Αλληλουχίας	Διαχωρίσιμο	CNC	Κλάση όχι κλικ
1-3	75.94%	76.03%	80.43%	74.28%
1-5	82.87%	82.92%	82.74%	81.88%
6-10	97.17%	97.18%	97.17%	97.18%

Η σύγκριση των 2 μοντέλων της βιβλιογραφίας συνεχίζει να είναι εντυπωσιακά κοντά. Το σημείο στο οποίο αξίζει να επικεντρωθούμε είναι στην απόδοση του CNC στις θέσεις 1-3. Στις αρχικές θέσεις το μοντέλο αυτό ξεπερνά τα μοντέλα της βιβλιογραφίας πάνω από 4 ποσοστιαίες μονάδες το οποίο αποτελεί μεγάλη διαφορά. Σε αυτό το σημείο αξίζει να αναφέρουμε ότι **οι θέσεις επιχορηγούμενης αναζήτησης είναι συνήθως λιγότερες από 5.** Το παραπάνω δίνει μεγαλύτερη βαρύτητα στην επιτυχία του μοντέλου αυτού αφού μας ενδιαφέρει η επιτυχία στις πρώτες θέσεις καθώς στις επιχορηγούμενες αναζητήσεις έχουμε λιγότερες θέσεις.

Κεφάλαιο 6

Επίλογος

6.1 Σύνοψη και συμπεράσματα

Το πρόβλημα των δημοπρασιών επιδοτούμενων διαφημίσεων είναι ένα πρόβλημα του οποίου η μελέτη είναι εξαιρετικά σημαντική για τους κολοσσούς του διαδικτύου. Κάθε απειροελάχιστη αλλαγή μπορεί να οδηγήσει σε διαφορετικές τιμές ισορροπίας στα παίγνια και επομένως αλλαγές δισεκατομμυρίων των εσόδων.

Η μελέτη της διπλωματικής αυτής δείχνει πως τα περιθώρια για αλλαγές είναι περιορισμένα. Λόγω των ιδιοτήτων του GSP, επιβάλλονται περιορισμοί για το μοντέλο πρόβλεψης της συμπεριφοράς του χρήστη που θα χρησιμοποιηθεί. Τα μοντέλα της βιβλιογραφίας καθώς και το μοντέλο που προτείνεται στην διπλωματική αυτή, βρίσκονται εξαιρετικά κοντά και διαφέρουν κατά πολύ λίγο.

Παρ΄ όλα αυτά, λόγω των μεγεθών, κάθε μικρή διαφορά έχει μεγάλη σημασία. Η διπλωματική αυτή καταλήγει πως το διαχωρίσιμο μοντέλο έχει την μεγαλύτερη ακρίβεια σε σχέση με τα διαφορετικά μοντέλα. Όμως επειδή ενδεχομένως σκοπός μας είναι να επικεντρωθούμε στην πρόβλεψη των κλικ, το μοντέλο Αλληλουχίας ή το προτεινόμενο από την διπλωματική αυτή αποδίδουν ισχυρότερα στον τομέα αυτό.

Το σημαντικότερο από τα συμπεράσματα που βγήκαν είναι πως το προτινόμενο μοντέλο από την διπλωματική αυτή αποδίδει εξαιρετικά καλύτερα στις πρώτες θέσεις της αναζήτησης. Αυτό το συμπέρασμα είναι ιδιαίτερης σημασίας καθώς στις επιχορηγούμενες αναζητήσεις μας ενδιαφέρουν οι πρώτες 3 με 5 θέσεις καθώς ο αριθμός των διαφημίσεων είναι πάντα περιορισμένος. Για παράδειγμα, η Google παρουσιάζει 4 διαφημίσεις στην μέση περίπτωση.

6.2 Μελλοντική εργασία

Υπάρχουν αρκετές κατευθύνσεις για μελλοντική εργασία :

- Εάν βρεθούν τα κατάλληλα δεδομένα, ενδιαφέρουσα θα είναι η ανάλυση με χρήση μεθόδων μηχανικής μάθησης. Όπως αναφέρθηκε στο υποκεφάλαιο 4.3, υπάρχουν μία σειρά από σενάρια στα οποία η μηχανική μάθηση θα υπερτερούσε σε σχέση με την ανάλυση με ιστογράμματα η οποία συνέβη στην διπλωματική αυτή.

- Ενδιαφέρουσα επίσης θα αποτελούσε η σύγκριση των οργανικών αποτελεσμάτων με τα αποτελέσματα των επιχορηγούμενων αποτελεσμάτων κατά την αναζήτηση στο διαδίκτυο από τον χρήστη ώστε να γίνει γνωστό αν ο χρήστης έχει την ίδια συμπεριφορά και στις 2 περιπτώσεις.

- Η μελέτη της ακρίβειας ανά θέση θα μπορούσε να γίνει πιο συστηματικά. Με αυτόν τον τρόπο μπορούσε να γίνει καλύτερη ανάλυση των αδυναμιών των μοντέλων καθώς και να δημιουργηθούν προβλέπτες που χρησιμοποιούν σε κάθε θέση διαφορετικό μοντέλο.

- Εξαιρετικά φιλόδοξη θα ήταν η προσπάθεια για την δημιουργία ενός μοντέλου το οποίο δεν είναι ακολουθιακό. Οι μη ακολουθιακές προσπάθειες που έγιναν σε αυτήν την διπλωματική απέτυχαν αλλά στοιχεία δείχνουν πως ο χρήστης συμπεριφέρεται μη ακολουθιακά πολλές φορές [19].

- Η μελέτη των θέσεων 5-10 θα μπορούσε να γίνει ξεχωριστά ώστε να δημιουργηθούν μοντέλα τα οποία εντοπίζουν τις εξαιρέσεις (μόνο το 2.83% αποτελεί κλικ στις θέσεις αυτές).

- Τέλος, με μεγάλη υπολογιστική δύναμη, ενδιαφέρουσα θα ήταν η βελτιστοποίηση των παραμέτρων. Αυτό συμβαίνει καθώς για παράδειγμα τα λ_i του διαχωρίσιμου μοντέλου μπορούν να υπολογιστούν πειραματικά και όχι να προκύψουν από την ανάλυση των προβολών κάθε σελίδας όπως συνέβη στην διπλωματική αυτή.

Bibliography

[1] M. R. Arbabshirani, B. K. Fornwalt, G. J. Mongelluzzo, J. D. Suever, B. D. Geise, A. A. Patel, and G. J. Moore. Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. *npj Digital Medicine*, 1(1), apr 2018.

[2] M. Arsenis. Algorithmic game theory introduction to mechanism design [pdf]. 2016.

[3] S. Athey and G. Ellison. Position auctions with consumer search. Working Paper 15253, National Bureau of Economic Research, August 2009.

[4] L. Bottou, J. Peters, J. Q. nonero Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14:3207–3260, 2013.

[5] F. Brandt, V. Conitzer, U. Endriss, J. Lang, and A. D. Procaccia. *Handbook of Computational Social Choice*. Cambridge University Press, New York, NY, USA, 1st edition, 2016.

[6] R. Cavallo, P. Krishnamurthy, M. Sviridenko, and C. A. Wilkens. Sponsored search auctions with rich ads. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 43–51, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.

[7] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. In *OSDI'04: Sixth Symposium on Operating System Design and Implementation*, pages 137–150, San Francisco, CA, 2004.

[8] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, Jan. 2008.

[9] U. o. U. Department of Mathematics. Using the receiver operating characteristic (roc) curve to analyze a classification model [pdf]. 2017.

[10] B. Edelman, M. Ostrovsky, and M. Schwarz. Internet advertising and the generalized second price auction: Selling billions of dollars worth of keywords. Working Paper 11765, National Bureau of Economic Research, November 2005.

[11] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June 2008.

[12] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, jun 2006.

[13] D. Fotakis. Approximate mechanism design without money [pdf]. 2013.

[14] D. Fotakis, P. Krysta, and O. Telelis. Externalities among advertisers in sponsored search. In G. Persiano, editor, *Algorithmic Game Theory*, pages 105–116, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[15] S. Ghemawat, H. Gobioff, and S.-T. Leung. The google file system. *ACM SIGOPS Operating Systems Review*, 37(5):29, dec 2003.

[16] I. Giotis and A. R. Karlin. On the equilibria and efficiency of the GSP mechanism in keyword auctions with externalities. In *Lecture Notes in Computer Science*, pages 629–638. Springer Berlin Heidelberg, 2008.

[17] R. Gomes, N. Immorlica, and E. Markakis. Externalities in keyword auctions: An empirical and theoretical assessment. In S. Leonardi, editor, *Internet and Network Economics*, pages 172–183, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[18] P. Hummel and R. P. McAfee. Position auctions with externalities and brand effects. *CoRR*, abs/1409.4687, 2014.

[19] P. Jeziorski and I. Segal. What makes them click: Empirical analysis of consumer demand for search advertising. *American Economic Journal: Microeconomics*, 7(3):24–53, aug 2015.

[20] D. Kempe and M. Mahdian. A cascade model for externalities in sponsored search. In C. Papadimitriou and S. Zhang, editors, *Internet and Network Economics*, pages 585–596, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[21] J. Levin. Sponsored search auctions [pdf], stanford university. 2009.

[22] V. A. Luis. Sponsored search (pdf). 15–396: Science of the web course notes. carnegie mellon university. retrieved 2015-04-13. 2013.

[23] O. Maimon and L. Rokach. *Data Mining and Knowledge Discovery Handbook*. Springer-Verlag, Berlin, Heidelberg, 2005.

[24] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani. *Algorithmic Game Theory*. Cambridge University Press, 2007.

[25] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll. An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, 96(1):3–14, 2002.

[26] D. M. W. Powers. Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

[27] T. Roughgarden. *Twenty Lectures on Algorithmic Game Theory*. Cambridge University Press, 2016.

[28] P. Shah, M. Yang, S. Alle, A. Ratnaparkhi, B. Shahshahani, and R. Chandra. A practical exploration system for search advertising. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 1625–1631, New York, NY, USA, 2017. ACM.

[29] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[30] F. Stamos. https://github.com/filippst/hadoop, 2018.

[31] T. White. *Hadoop: The Definitive Guide*. O'Reilly Media, Inc., 4th edition, 2015.

[32] C. Willmott and K. Matsuura. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30:79–82, 2005.

[33] X. Zhang, N. Quadrianto, K. Kersting, Z. Xu, Y. Engel, C. Sammut, M. Reid, B. Liu, G. I. Webb, C. Sammut, M. Sipper, L. Saitta, M. Sebag, C. C. Aggarwal, T. Gärtner, T. Horváth, S. Wrobel, D. Chakrabarti, J. McAuley, T. Caetano, W. Buntine, T. R. Jensen, C. Sammut, L. Holder, H. Sharara, and L. Getoor. Genetic and evolutionary algorithms. In *Encyclopedia of Machine Learning*, pages 456–457. Springer US, 2011.

Γλωσσάριο

Ελληνικός όρος	Αγγλικός όρος
ακρίβεια	accuracy
γενικευμένη δημοπρασία δεύτερης τιμής	generalized second price auction
δημοπρασία	auction
δημοπρασίες επιχορηγούμενης αναζήτησης	sponsored search auctions
διαχωρίσιμο μοντέλο	separable model
εμβαδόν κάτω απ΄ την καμπύλη	area under the curve
θεωρία παιγνίων	game theory
μοντέλο αλληλουχίας	cascade model
κοινωνική επιλογή	social choice
λογιστική παλινδρόμηση	logistic regression
μηχανική μάθηση	machine learning
πιθανότητα συνέχειας	continuation probability
πίνακας σύγχυσης	confusion matrix
συνεδρία	session
σχεδιασμός μηχανισμών	mechanism design
χαρακτηριστικό	feature