ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Διπλωματική Εργασία

της φοιτήτριας του Τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών της Πολυτεχνικής Σχολής του Πανεπιστημίου Θεσσαλίας

ΖΩΓΟΠΟΥΛΟΥ ΑΝΝΑΣ του ΚΩΝΣΤΑΝΤΙΝΟΥ

Αριθμός Μητρώου: 835

Θέμα

ΑΛΓΟΡΙΘΜΟΙ ΔΙΑΦΗΜΙΣΗΣ ΣΤΟΝ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ

Επιβλέποντες

Μποζάνης Παναγιώτης, Αναπληρωτής Καθηγητής

Τσομπανοπούλου Παναγιώτα, Επίκουρος Καθηγήτρια

Βόλος, Ιούνιος 2014

Θέμα: ΑΛΓΟΡΙΘΜΟΙ ΔΙΑΦΗΜΙΣΗΣ ΣΤΟΝ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ

Φοιτήτρια: Επιβλέποντες:

Ζωγοπούλου Άννα Μποζάνης Παναγιώτης, Αναπληρωτής Καθηγητής

Τσομπανοπούλου Παναγιώτα, Επίκουρος Καθηγήτρια

Περίληψη

Η διαφήμιση στον παγκόσμιο ιστό είναι μια επιχείρηση προσοδοφόρα και αποτελεί την κύρια πηγή εσόδων των δημοφιλέστερων ιστοσελίδων, όπως είναι η Google, η Yahoo και το Facebook. Οι διαφημιστές κερδίζουν τους πελάτες τους κυρίως μέσα από δύο τρόπους: από τη διαδικασία του *sponsored search* (αναζήτηση με χορηγία), που είναι η εμφάνιση διαφημίσεων κειμένου δίπλα στα αποτελέσματα της μηχανής αναζήτησης και από τη διαδικασία του *display advertising* (διαφήμιση στην οθόνη), που είναι η εμφάνιση διαφημίσεων σε γραφική μορφή ως μέρος μιας ιστοσελίδας. Στην πιο απλή μορφή της διαφήμισης στον παγκόσμιο ιστό, παίρνουν μέρος τρείς οντότητες: οι διαφημιστές, το μέσο και οι χρήστες. Οι δράσεις των οντοτήτων αυτών μπορούν να συνοψιστούν σε τρία βήματα: το *bidding*, που είναι η δράση μεταξύ διαφημιστή και μέσου, το *delivery*, που είναι η δράση μεταξύ μέσου και χρηστών και τέλος, το *response*, που είναι η ανταπόκριση του χρήστη και συνεπώς η δράση μεταξύ χρηστών και διαφημιστή. Σε αυτήν την εργασία μελετάμε καθένα από τα βήματα αυτά σε ξεχωριστό κεφάλαιο και περιγράφουμε αλγόριθμους και στρατηγικές που έχουν παρουσιαστεί στη βιβλιογραφία και σε διεθνή συνέδρια τα τελευταία τρία χρόνια.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον Αναπληρωτή Καθηγητή κ. Παναγιώτη Μποζάνη και την Επίκουρο Καθηγήτρια κα. Παναγιώτα Τσομπανοπούλου, επιβλέποντες της διπλωματικής εργασίας, που με συμβούλευαν και με καθοδηγούσαν για τη διεκπεραίωσή της. Η συνεργασία μας ήταν άψογη.

Ένα μεγάλο ευχαριστώ στους φίλους μου, που ήταν ένα αναπόσπαστο κομμάτι της ζωής μου στο Βόλο. Οι στιγμές που περάσαμε μαζί θα μου μείνουν αξέχαστες πραγματικά. Αισθάνομαι ιδιαίτερα χαρούμενη και πολύ τυχερή που τους γνώρισα.

Τέλος, η οικογένεια μου, με την αγάπη και τη στήριξη της, μου έδινε πάντα θάρρος για να προχωράω. Τους ευχαριστώ όλους!

Ζωγοπούλου Άννα

Περιεχόμενα

Κατάλογος σχημάτων

Κεφάλαιο 1: Εισαγωγικά – Βασικές έννοιες

Οι διαφημίσεις στον παγκόσμιο ιστό είναι ένα αναπόσπαστο κομμάτι της σύγχρονης πραγματικότητας. Χρησιμοποιούνται προκειμένου να προωθήσουν ένα προϊόν ή μια υπηρεσία στους καταναλωτές, οδηγώντας τους είτε σε μια τοποθεσία αγοράς, είτε ακόμη και σε μια άλλη ιστοσελίδα. Από τις αρχές τις δεκαετίας του 1990 παρατηρείται μια τεράστια αύξηση στον τομέα των διαφημίσεων στον παγκόσμιο ιστό, η οποία συνεχίζεται με ταχείς ρυθμούς μέχρι και σήμερα. Το Διαδίκτυο, ως ένα παγκόσμιο μέσο επικοινωνίας, παρέχει στους διαφημιζόμενους μοναδική και συχνά οικονομικότερη και αποτελεσματικότερη εμπειρία. Οι καταναλωτές χρησιμοποιούν το Διαδίκτυο για περισσότερα πράγματα από απλή ψυχαγωγία, όπως για παράδειγμα κάνουν με το ραδιόφωνο, την τηλεόραση, τα περιοδικά και τις εφημερίδες. Χρησιμοποιούν το Διαδίκτυο για να τους βοηθήσει σχεδόν σε κάθε πτυχή της ζωής τους. Έτσι δημιουργούνται αμέτρητες ευκαιρίες για να τοποθετηθούν στοχευμένα μηνύματα διαφήμισης.

Αναφερόμαστε δηλαδή, σε διαφημίσεις ηλεκτρονικού ταχυδρομείου, διαφημίσεις μηχανών αναζήτησης, μάρκετινγκ μέσα από τα μέσα μαζικής ενημέρωσης και τα κοινωνικά δίκτυα, αλλά και διαφημίσεις μέσω του κινητού τηλεφώνου, όπως και διάφοροι άλλοι τύποι διαφημίσεων προβολής(π.χ. Web banner advertising[1]). Οι διαφημίσεις αυτές περιλαμβάνουν συνήθως έναν εκδότη, ο οποίος ενσωματώνει τις διαφημίσεις σε απευθείας σύνδεση με το περιεχόμενό τους, και έναν διαφημιστή, ο οποίος παρέχει τις διαφημίσεις που θα εμφανίζονται με το περιεχόμενο του εκδότη. Άλλοι πιθανοί συμμετέχοντες της διαδικασίας είναι διάφορες διαφημιστικές εταιρίες που συμβάλλουν στη δημιουργία και τοποθέτηση αγγελιών ή ένας διακομιστής διαφημίσεων (ad server), που προσφέρει τεχνολογική υποστήριξη για τα στατιστικά στοιχεία μιας διαφήμισης, αλλά και τέλος μια θυγατρική διαφημιστική εταιρία, η οποία εκτελεί ανεξάρτητες διαφημιστικές εργασίες για τον διαφημιζόμενο. Ως αποτέλεσμα όλων των παραπάνω, κρίνεται αναγκαίος ο σχεδιασμός αλγορίθμων και στρατηγικών που θα κατατάσσουν τις καλύτερες διαφημίσεις προς επιλογή, αλλά και θα μετρούν την επιρροή των διαφημίσεων στους χρήστες.

Μέχρι στιγμής η πιο προσοδοφόρα επιλογή για την online διαφήμιση απαιτεί αναζήτηση και γενικά γίνεται προσπάθεια ταιριάσματος αιτημάτων αναζήτησης με διαφημίσεις. Ως εκ τούτου, αφιερώνεται ιδιαίτερη προσπάθεια στην ανεύρεση αλγορίθμων για την βελτιστοποίηση του τρόπου που γίνεται αυτή η ανάθεση. Αυτό λοιπόν, είναι μια κατηγορία αλγορίθμων την οποία θα μελετήσουμε στη συνέχεια.

Να σημειωθεί ότι η μελέτη μας γίνεται πάνω σε online αλγόριθμους. Αυτό σημαίνει ότι υπάρχουν φορές που δεν μπορούμε να δούμε όλα τα δεδομένα πριν ο αλγόριθμός μας πάρει κάποιες αποφάσεις, κάτι που έρχεται σε αντιδιαστολή με τους τυπικούς αλγόριθμους. Αυτοί, οι λεγόμενοι και offline αλγόριθμοι λειτουργούν ως εξής: όλα τα δεδομένα που απαιτούνται από τον αλγόριθμο παρουσιάζονται αρχικά. Ο αλγόριθμος μπορεί να έχει πρόσβαση στα δεδομένα με οποιαδήποτε σειρά. Στο τέλος, ο αλγόριθμος παράγει το αποτέλεσμα του.

[1] Αυτή η μορφή της online διαφήμισης συνεπάγεται την ενσωμάτωση μιας διαφήμισης σε μια ιστοσελίδα

Συνοψίζοντας, θα περιγραφούν και θα κατηγοριοποιηθούν στη συνέχεια αλγόριθμοι και στρατηγικές της online διαφήμισης στον παγκόσμιο ιστό από μελέτη σε βιβλιογραφία και δημοσιευμένες μελέτες συνεδρίων των τριών τελευταίων χρόνων.

1.1 Επισκόπηση των διαφημίσεων στον παγκόσμιο ιστό

1.1.1 Κατηγοριοποίηση διαφημίσεων

Διακρίνουμε δύο μεγάλες κατηγορίες διαφημίσεων: **Sponsored search** (αναζήτηση με χορηγία) που είναι η εμφάνιση διαφημίσεων κειμένου δίπλα στα αποτελέσματα της μηχανής αναζήτησης, και **Display advertising** (διαφήμιση στην οθόνη) που είναι η εμφάνιση διαφημίσεων σε γραφική μορφή ως μέρος μιας ιστοσελίδας.

Πιο συγκεκριμένα, στο sponsored search έχουμε να κάνουμε με τις επονομαζόμενες και ως διαφημίσεις αναζήτησης, στις οποίες ο διαφημιστής μπορεί να συμπεριλαμβάνεται στα αποτελέσματα μιας αναζήτησης για επιλεγμένες λέξεις κλειδιά. Οι διαφημίσεις αυτές πωλούνται συχνά μέσω δημοπρασιών σε πραγματικό χρόνο, όπου οι διαφημιστές υποβάλουν κάποια προσφορά για τις λέξεις κλειδιά. Εκτός από τον καθορισμό μιας ανώτατης τιμής ανά λέξη κλειδί, οι προσφορές μπορεί να περιλαμβάνουν και κάποιες επιπλέον ιδιότητες όπως το χρόνο, τη γλώσσα, τη γεωγραφική περιοχή καθώς και άλλους περιορισμούς. Οι μηχανές αναζήτησης τοποθετούν κατά σειρά τις υψηλότερες σε προσφορά διαφημίσεις, αλλά λαμβάνουν υπόψη τους και άλλους παράγοντες, όπως την ποιότητα της ιστοσελίδας, τη σχετικότητα της λέξης κλειδί με τη διαφήμιση.

Όσον αφορά το display advertising, έχουμε τη διαδικασία όπου το διαφημιστικό μήνυμα παρουσιάζεται οπτικά χρησιμοποιώντας κείμενο, λογότυπα, κινούμενα σχέδια, βίντεο, φωτογραφίες ή άλλα γραφικά. Απευθύνεται συχνά σε χρήστες με συγκεκριμένα χαρακτηριστικά προκειμένου να αυξηθεί η επιρροή των διαφημίσεων στους χρήστες. Εδώ συναντάμε και τον όρο «cookies», που είναι ουσιαστικά κάποια μοναδικά αναγνωριστικά υπολογιστών, για να αποφασιστεί ποιές διαφημίσεις θα εξυπηρετούσαν έναν συγκεκριμένο καταναλωτή. Επίσης, μπορούν να παρακολουθούν αν ένας χρήστης φύγει από μια σελίδα χωρίς να αγοράσει τίποτα, έτσι ώστε ο διαφημιστής να μπορεί αργότερα να στοχεύει σε διαφημίσεις από την ιστοσελίδα που ο χρήστης επισκέφθηκε.

Οι διαφημιστές επίσης, συλλέγουν δεδομένα σχετικά με τη δραστηριότητα του χρήστη από πολλούς εξωτερικούς δικτυακούς τόπους, για να μπορούν να δημιουργήσουν μια λεπτομερή εικόνα του χρήστη και έτσι να προσφέρουν πιο στοχευμένα. Αυτή η ομαδοποίηση των δεδομένων ονομάζεται στόχευση συμπεριφοράς (behavioral targeting). Η επαναστόχευση, η συμπεριφορική στόχευση, γενικά όλα έχουν σχεδιαστεί για να αυξήσουν την απόδοση ενός διαφημιστή. Οι διαφημιστές μπορούν τέλος να παραδώσουν διαφημίσεις με βάση τη γεωγραφία ενός χρήστη μέσω γεωγραφικής στόχευσης. Η διεύθυνση IP ενός χρήστη δίνει κάποιες γεωγραφικές πληροφορίες και μπορεί να συμπληρωθούν και να τελειοποιηθούν με άλλες πληροφορίες και έτσι να περιοριστεί το εύρος των πιθανών θέσεων του χρήστη. Για παράδειγμα, με φορητές συσκευές, οι διαφημιστές μπορούν να χρησιμοποιούν τον δέκτη GPS ενός τηλεφώνου, τα «cookies» και

άλλα σταθερά δεδομένα στον υπολογιστή ενός χρήστη και έτσι να στενεύει περαιτέρω η θέση που μπορεί να βρίσκεται ένας χρήστης.

1.1.2 Βήματα διάκρισης

Σε δεύτερη φάση, στις διαφημίσεις του παγκόσμιου ιστού δραστηριοποιούνται τρείς βασικές οντότητες: οι διαφημιστές, οι χρήστες και το διαφημιστικό μέσο, όπως είναι οι μηχανές αναζήτησης για το sponsored search ή οι εκδότες των ιστοσελίδων για το display advertising. Ανάμεσα σε αυτούς λοιπόν, μπορούμε να διακρίνουμε τρία βήματα: το **bidding**, που είναι η δράση μεταξύ διαφημιστή και μέσου, το **delivery**, που είναι η δράση μεταξύ μέσου και χρηστών και τέλος, το **response**, που είναι η ανταπόκριση του χρήστη και συνεπώς δράση μεταξύ χρηστών και διαφημιστή.

Σε μεγαλύτερη ανάλυση στο βήμα του bidding, δηλαδή όταν έχουμε την υποβολή προσφορών, έχουμε την αλληλεπίδραση μέσου και διαφημιστή. Ο διαφημιστής παρέχει στο μέσο τις διαφημίσεις του, τις προτιμήσεις του για την παράδοσή τους, καθώς και την τιμή που είναι διατεθειμένος να πληρώσει. Οι εκφράσεις προτίμησης (bid phrases) στο sponsored search εκφράζονται μέσα από τις λέξεις κλειδιά (keywords). Ένας διαφημιστής θέλει η διαφήμισή του να εμφανίζεται σε ένα χρήστη, του οποίου το ερώτημα (query) ταιριάζει με μία από τις λέξεις κλειδιά που παρέχονται. Στο display advertising, ο διαφημιστής επιλέγει μια σειρά από ιστοσελίδες και ένα χρονοδιάγραμμα, γιατί θέλει οι διαφημίσεις του να εμφανίζονται όταν ένας χρήστης επισκέπτεται τις παρεχόμενες σελίδες κατά τη διάρκεια του καθορισμένου χρονοδιαγράμματος. Επίσης, ο διαφημιστής μπορεί να εκφράσει τις προτιμήσεις του όσον αφορά τα δημογραφικά χαρακτηριστικά, π.χ. το φύλο ή την ηλικία, από τους χρήστες που βλέπουν τις διαφημίσεις του. Όσον αφορά την τιμολόγηση και τις πληρωμές στο sponsored search, η μηχανή αναζήτησης επιβαρύνει χρηματικά τον διαφημιστή για κάθε κλικ στις διαφημίσεις. Η τιμή ενός κλικ δεν μπορεί να υπερβαίνει την προσφορά του διαφημιστή, δηλαδή το μέγιστο ποσό που είναι διατεθειμένος να πληρώσει. Στο display advertising, η ιστοσελίδα του εκδότη συνήθως χρεώνει τον διαφημιστή για κάθε εμφάνιση των διαφημίσεων στην ιστοσελίδα του εκδότη. Η τιμή για κάθε εμφάνιση και ο συνολικός αριθμός των εμφανίσεων πάνω από ένα ορισμένο χρονικό διάστημα, έχουν συμφωνηθεί σε μια σύμβαση μεταξύ του εκδότη και του διαφημιστή.

Το βήμα του delivery (παράδοση), αναφέρεται στην αλληλεπίδραση μεταξύ του μέσου και των χρηστών. Ως αποτέλεσμα του σταδίου υποβολής προσφορών (bidding), το μέσο έχει ανά πάσα στιγμή μια καταγραφή των αγγελιών. Όταν ένας χρήστης επισκέπτεται μια ιστοσελίδα, το μέσο θα πρέπει να επιλέξει μία από αυτές τις διαφημίσεις και να την εμφανίσει στο χρήστη. Το πιο δύσκολο πρόβλημα στο στάδιο αυτό είναι η επιλογή διαφημίσεων, και αυτό γιατί πρέπει να συμμορφώνονται με τις προτιμήσεις του διαφημιζόμενου, αλλά θα πρέπει και να βελτιστοποιούν τη χρήση του καταλόγου από την πλευρά του μέσου. Το μέσο συνήθως επιλέγει τις διαφημίσεις σε μια διαδικασία δύο σταδίων. Στο πρώτο στάδιο, το μέσο καθορίζει τις επιλέξιμες διαφημίσεις, δηλαδή, τις διαφημίσεις που συμμορφώνονται με τις προτιμήσεις του διαφημιστή. Η επιλογή γίνεται σε όσο το δυνατόν πραγματικό χρόνο και το μέσο θα χρειαστεί να διατηρεί δείκτες που θα διευκολύνουν τη γρήγορη ανάκτηση των διαφημίσεων που πληρούν ορισμένες

προϋποθέσεις. Σε δεύτερο στάδιο, το μέσο καθορίζει ποια από τις ενεργές διαφημίσεις να δείξει στο χρήστη μέσω μιας δημοπρασίας μεταξύ των επιλεγμένων διαφημίσεων.

Το τρίτο βήμα, δηλαδή αυτό του response (απόκριση), αναφέρεται στην απόκριση των χρηστών στις διαφημίσεις. Ένας χρήστης που βλέπει μια διαφήμιση μπορεί είτε να την αγνοήσει, είτε να προβεί σε ενέργειες που σχετίζονται με τη διαφήμιση. Για παράδειγμα, ο χρήστης μπορεί να κάνει κλικ στη διαφήμιση για να επισκεφθεί τη σελίδα του διαφημιστή. Στην πράξη, η μετρούμενη ανταπόκριση του χρήστη σε μια διαφήμιση περιορίζεται συνήθως στο αν τελικά κάνει κλικ στη διαφήμιση.

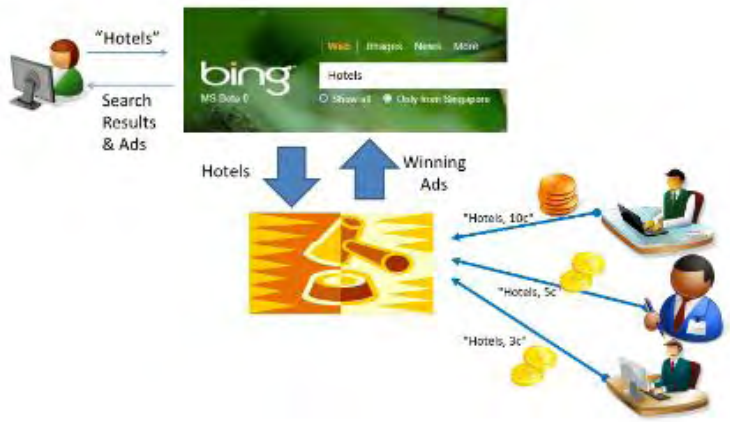
1.1.3 Γενικές έννοιες και ορολογία

Στο σημείο αυτό θα γίνει μια αναφορά σε κάποιους χρήσιμους όρους που συναντώνται στη συνέχεια της εργασίας, αλλά θα αναφερθούν και κάποια επιπλέον στοιχεία για κάθε κατηγορία από αυτές που περιγράφηκαν στις προηγούμενες παραγράφους (1.1.1, 1.1.2).

Όταν ένας χρήστης θέτει ένα ερώτημα (**query**) στη μηχανή αναζήτησης γίνονται δύο ξεχωριστές αναζητήσεις: **web search** που είναι η αναζήτηση στο σώμα των ιστοσελίδων και **sponsored search** που πρόκειται για διαφημίσεις που εμφανίζονται επάνω και στα πλάγια των αποτελεσμάτων αναζήτησης. Μέσω του sponsored search υπολογίζεται ένα σύνολο διαφημίσεων που προωθούν προϊόντα και υπηρεσίες. Συνηθίζεται στον κόσμο των διαφημίσεων το κείμενο της διαφήμισης να είναι ορατό στον χρήστη (**ad creative**) και να παράγεται από τον διαφημιστή για να μεγιστοποιήσει το ενδιαφερόμενο κοινό. Η επιλογή των διαφημίσεων εξαρτάται από τις λεγόμενες λέξεις προσφοράς (**bid phrases**), δηλαδή τη φράση που ο διαφημιστής έχει ορίσει ως ταιριαστή για την συγκεκριμένη διαφήμιση. Βέβαια, είναι αρκετά δύσκολο ο διαφημιστής να βρει όλα τα αιτήματα που να είναι σχετικά με τις διαφημίσεις του. Έτσι, οι μηχανές αναζήτησης επιτρέπουν και προηγμένη (advanced) αντιστοίχηση αιτήματος με διαφήμιση, δηλαδή μια διαφήμιση μπορεί να επιλεγεί ακόμα και αν η φράση προσφοράς (bid phrase) δεν ταιριάζει με το αίτημα.

Οι διαφημιστές επίσης, τοποθετούν προσφορές σε μια φράση που αποτελεί το αντικείμενο αναζήτησης, και οι θέσεις τους στη στήλη διαφημίσεων στη σελίδα των αποτελεσμάτων της μηχανής αναζήτησης καθορίζεται από την προσφορά. Έτσι, κάθε διαφήμιση κουβαλά μία ή περισσότερες εκφράσεις προσφοράς (bid phrases). Επιπλέον, μια διαφήμιση περιέχει έναν τίτλο (**title**) συνήθως να εμφανίζεται με έντονη γραφή. Επίσης, συνοδεύεται από το λεγόμενο στα αγγλικά **creative**, που είναι λίγες γραμμές κειμένου που εμφανίζονται στη σελίδα. Τέλος, γίνεται αναφορά στον όρο **landing page**, που είναι ουσιαστικά η σελίδα της διαφήμισης που ο χρήστης οδηγείται μέσω του URL που περιέχεται στη διαφήμισή του.

1.2 Το σύστημα sponsored search



Σχήμα 1 : Η διαδικασία του sponsored search [Πηγή: από Βιβλ. Κεφ., αναφορά 2]

Στην εικόνα παρουσιάζεται συνοπτικά η διαδικασία του sponsored search. Όταν ένας χρήστης ψάχνει για τη λέξη «hotels», η μηχανή αναζήτησης στέλνει τη λέξη κλειδί από αυτό το αίτημα του χρήστη για μια δημοπρασία ανάμεσα στους διαφημιστές, οι οποίοι διατίθενται να πληρώσουν συγκεκριμένο ποσό χρημάτων για κάθε κλικ του χρήστη.

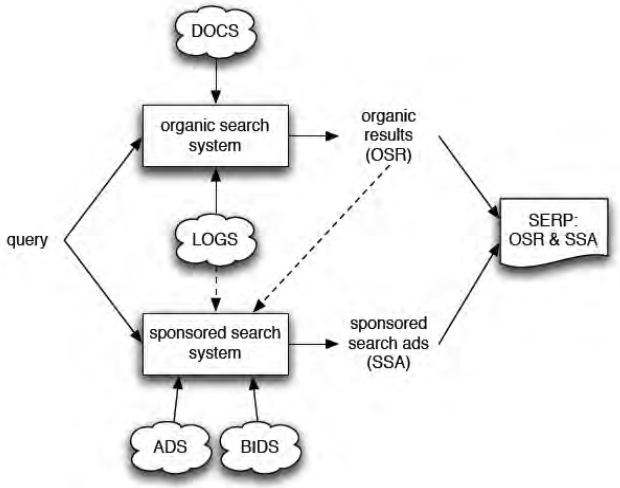
Η διαδικασία του sponsored search λοιπόν, αποτελείται από τρεις οντότητες:

α) τον **διαφημιστή** (advertiser), που παρέχει τις διαφημίσεις

β) τη **μηχανή αναζήτησης** (search engine), που παρέχει το χώρο που θα τοποθετηθούν οι διαφημίσεις και επιλέγει διαφημίσεις που σχετίζονται με το αίτημα του χρήστη

γ) τους **χρήστες**, οι οποίοι επισκέπτονται μια ιστοσελίδα και αλληλεπιδρούν με τις διαφημίσεις.

Στο παρακάτω σχήμα φαίνεται η αρχιτεκτονική του συστήματος που αφορά την κατηγορία sponsored search σε γενική μορφή.

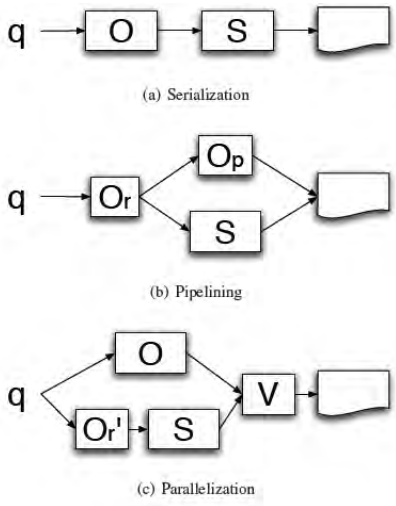


Σχήμα 2 : Η αρχιτεκτονική του συστήματος Sponsored Search [Πηγή: από Βιβλ. Κεφ., αναφορά 1]

SERP: search engine results page (η σελίδα αποτελεσμάτων της μηχανής αναζήτησης)

Ακολουθεί μια μικρή επεξήγηση του παραπάνω σχήματος (Σχήμα 2). Ένα αίτημα του χρήστη q χωρίζεται σε δύο διαφορετικά στοιχεία τα αποτελέσματα των οποίων ενώνονται στη σελίδα των συνολικών αποτελεσμάτων της μηχανής αναζήτησης. Η καθυστέρηση δίνεται από τον τύπο $l = \max(t_o,t_s)$, όπου το t_o είναι ο χρόνος που απαιτείται για αναζήτηση τύπου organic και t_s είναι ο απαιτούμενος χρόνος για το στοιχείο τύπου sponsored search (βλέπε σχήμα 2).

Στη συνέχεια παρουσιάζονται τρεις διαφορετικές προσεγγίσεις του συστήματος sponsored search: α) σειριακά β) σε μορφή pipeline γ) παράλληλα. Έτσι έχουμε το παρακάτω σχήμα:



Σχήμα 3 : Οι τρείς προσεγγίσεις του συστήματος Sponsored Search [Πηγή: από Βιβλ. Κεφ., αναφορά 1].

Επεξηγείται με λίγα λόγια το παραπάνω σχήμα (Σχήμα 3).

α) Η καθυστέρηση δίνεται από τον τύπο $l = t_o + t_s$. Η έξοδος του στοιχείου τύπου organic περιέχει πληροφορία που δεν χρησιμοποιείται από το στοιχείο sponsored search. Αυτό το στοιχείο ανακτά όλες τις σελίδες που σχετίζονται με μια αίτηση του χρήστη και τις κατατάσσει. Τέλος, εκτελεί μια διαδικασία που έχει να κάνει με την επιλογή ανεξάρτητων αιτήσεων χρήστη, όπου συνοψίζει αποσπάσματα για κάθε σελίδα αποτελεσμάτων καθώς και αποτελεσμάτων έπειτα από διαδικασία ομαδοποίησης (σχήμα 3α).

β) 'Εχουμε τη δυνατότητα να χωρίσουμε το στοιχείο που εκτελεί την αναζήτηση σε δύο μέρη. Στο στοιχείο που είναι υπεύθυνο για ανάκτηση και κατάταξη (retrieval-ranking, Or) και στο στοιχείο που εκτελεί την μετά-επεξεργασία έργου(post-processing, Op). Η καθυστέρηση δίνεται από τον τύπο $l = t_r + \max(t_p, t_s)$, όπου το t_r είναι ο χρόνος που

απαιτείται για το στοιχείο Or και t_p ο χρόνος για το στοιχείο Op. Όμως, το Op παίρνει περισσότερο χρόνο από το Or και έτσι max (t_p, t_s) = t_p (σχήμα 3b).

γ) Η προηγούμενη περιγραφή, δηλαδή αυτή του pipeline(β), απαιτεί επικοινωνία ανάμεσα στα δύο στοιχεία, το Or και Op και άρα πρέπει να φροντίσουμε για το κόστος αυτό. Ένας τρόπος αντιμετώπισης είναι να δημιουργήσουμε ένα αντίγραφο του Or, το Or' και να το τοποθετήσουμε σαν κομμάτι του στοιχείου sponsored search (βλέπε σχήμα 3c). Η καθυστέρηση l θα είναι η ίδια με αυτή της προσέγγισης pipeline. Ακόμη, μπορούμε να μειώσουμε τον απαιτούμενο χώρο αποθήκευσης, καθώς και το χρόνο ανάκτησης ενός εγγράφου στο στοιχείο sponsored search, αν μειώσουμε το μέγεθος του καταλόγου που τα διατηρεί. Αυτό, έχει σαν αποτέλεσμα να μειώνονται τα έγγραφα που πρέπει να αντιγραφούν και τελικά να μειώνεται και ο χρόνος τρεξίματος της όλης διαδικασίας t_r'. Άρα, με δεδομένη την αίτηση q ενός χρήστη, το Or' θα δεικτοδοτήσει έγγραφα για τα οποία υπάρχει έστω μία σχετική διαφήμιση. Μετά, το στοιχείο S θα ανακτήσει τις διαφημίσεις που ταιριάζουν με την πιο υψηλή βαθμολογία σε κατάταξη. Τέλος, το στοιχείο V χρειάζεται, διότι απορρίπτει όποιες διαφημίσεις έρχονται από το S που αντιστοιχούν σε έγγραφα που δεν εμφανίζονται με υψηλό σκορ κατάταξης του O. Το στοιχείο V χρειάζεται χρόνο t_v για να αξιολογήσει τις διαφημίσεις. Στην περίπτωση αυτή η καθυστέρηση δίνεται από τον τύπο $l = \max(t_o, t_r' + t_s) + t_v$.

1.3 Οργάνωση επόμενων κεφαλαίων

Με βάση τα παραπάνω, από μελέτη σε βιβλιογραφία, δημοσιευμένες μελέτες συνεδρίων των τριών τελευταίων χρόνων (2010-2013), γίνεται η εξής κατηγοριοποίηση των αλγορίθμων και των στρατηγικών που υπάρχουν στον τομέα των διαφημίσεων στον παγκόσμιο ιστό:

Στο Κεφάλαιο 2, θα περιγραφούν και θα δοθούν χαρακτηριστικά παραδείγματα αλγορίθμων και στρατηγικών που αφορούν την γενικότερη κατηγορία Sponsored Search and Biding. Με άλλα λόγια, αναφερόμαστε στην ομάδα αλγορίθμων για διαφημίσεις που εμφανίζονται επάνω και στα πλάγια των αποτελεσμάτων αναζήτησης, σε συνδυασμό με το βήμα του biding, που είναι η δράση μεταξύ διαφημιστή και μέσου, όπως περιγράφτηκε αναλυτικά σε προηγούμενη παράγραφο (1.1.2). Πιο αναλυτικά, περιγράφεται η διαδικασία του Output URL Biding και συγκεκριμένα παρουσιάζεται ο αλγόριθμος για το πρόβλημα query set output cover. Λεπτομέρειες θα δοθούν στη συνέχεια. Επίσης, περιγράφεται και η διαδικασία της αυτόματης παραγωγής φράσεων προσφοράς για διαφημίσεις στον παγκόσμιο ιστό (Automatic generation of bid phrases for online advertising). Εδώ παρουσιάζονται απλά κάποια βασικά μοντέλα και στρατηγικές που αφορούν την διαδικασία αυτή, όπως το μοντέλο μετάφρασης ή το μοντέλο γλώσσας.

Στο Κεφάλαιο 3, περιγράφεται η γενικότερη κατηγορία Sponsored Search and Delivery, δηλαδή εστιάζουμε σε αλγορίθμους και στρατηγικές που αφορούν τη δράση μεταξύ μέσου και χρηστών (delivery). Πιο συγκεκριμένα, η πρώτη διαδικασία που μελετάται είναι το λεγόμενο SSA (Sponsored Search Auctions with Conflicts) και ο αλγόριθμος του Debbassac. Στη συνέχεια περιγράφονται κάποιοι αλγόριθμοι που έχουν να κάνουν με το πώς θα

ξοδεύονται όσο το δυνατόν λιγότερα χρήματα στη διαδικασία του search advertising (Optimizing budget constrained spend in search advertising) και χαρακτηριστικά περιγράφουμε τα βήματα του αλγόριθμου VPT (Vanilla Probabilistic Throttling). Σε επόμενο στάδιο, περιγράφονται τέσσερις αλγόριθμοι που ανήκουν στη γενικότερη κατηγορία της δομής μιας διαφήμισης, στη δομημένη δεικτοδότηση, αλλά και στην ανακάλυψη διαφημίσεων στη διαδικασία του sponsored search (The anatomy of an Ad: Structured indexing and retrieval for sponsored search).

Στο κεφάλαιο 4, αναλύεται η κατηγορία Sponsored Search and Response, δηλαδή εστιάζουμε σε αλγόριθμους και μοντέλα που αφορούν την ανταπόκριση του χρήστη και συνεπώς μιλάμε για τη δράση μεταξύ χρηστών και διαφημιστή. Το πρώτο μοντέλο που μελετάται είναι το GCM (General Click Model) και είναι ικανό να προβλέπει το πώς συμπεριφέρεται ο χρήστης στα URLs που εμφανίζονται στην μηχανή αναζήτησης. Σε δεύτερο στάδιο, περιγράφεται ένα ιεραρχικό μοντέλο για των υπολογισμό των τιμών μιας διαφήμισης και στη συνέχεια αναλύεται η διαδικασία της πρόβλεψης εμφάνισης των διαφημίσεων, με χρήση του δικτύου Bayes και του δυναμικού γραμμικού μοντέλου (DLM). Στο τέλος, γίνεται προσπάθεια πρόβλεψης της τιμής του CTR (Click Through Rate) με χρήση υβριδικών μοντέλων και περιγράφονται αλγόριθμοι για την εξόρυξη πληροφορίας για τη συμπεριφορά του χρήστη.

Στο κεφάλαιο 5, ασχολούμαστε με τη διαδικασία του Display Advertising and Delivery, δηλαδή τη δράση μεταξύ μέσου και χρηστών. Γίνεται μια εννοιολογική προσέγγιση της διαφήμισης αυτής και προσπαθούμε με διάφορες μεθόδους, όπως το σημασιολογικό – συντακτικό ταίριασμα να βρεθούν οι διαφημίσεις που ταιριάζουν καλύτερα στις προτιμήσεις του χρήστη. Στη συνέχεια, μελετούνται παράγοντες όπως το CTR (Click Through Rate), το CVR (Conversion Rate) για να καθοριστεί όσο το δυνατόν το πρόβλημα πρόβλεψης για μη εγγυημένη παράδοση μιας διαφήμισης. Τέλος, γίνεται αναφορά σε μεθόδους που χρησιμοποιούνται για το πρόβλημα της μεγιστοποίησης εσόδων, προσδίνοντας τιμές σε αυτά που παρουσιάζονται στην οθόνη (display ads).

Βιβλιογραφία κεφαλαίου

[1] Panagiotis Papadimitriou. “Algorithms and Strategies for Web Advertising”. Stanford University, Dept. of Electrical Engineering. December 2011

[2] Abhirup Nath, Shibnath Mukherjee, Prateek Jain, Navin Goyal, Srivatsan Laxman. “Ad Impression Forecasting for Sponsored Search”. In *Proceedings of the 22nd international conference on World Wide Web (WWW 2013).* Rio de Janeiro, Brazil. May 13-17, 2013.

[3] Anand Rajaraman, Jure Leskovec, Jeffrey D. Ullman. “Mining of Massive Datasets” (Chapter 8). Stanford University. 2013.

[4] http://en.wikipedia.org/wiki/Online_advertising

[5] http://www.techopedia.com/definition/26362/online-advertising

[6] http://www.entrepreneur.com/encyclopedia/online-advertising

[7] http://en.wikipedia.org/wiki/Web_banner

Κεφάλαιο 2: Sponsored Search & Bidding

2.1 Output URL bidding

Το Output URL bidding πρόκειται για έναν μηχανισμό όπου οι διαφημιστές τοποθετούν τις προσφορές τους στα URL των αποτελεσμάτων μιας αναζήτησης. Τα URLs είναι ένας τρόπος ο διαφημιστής να εκφράσει το ενδιαφέρον του για κάτι και οι μηχανές αναζήτησης να αναγνωρίσουν χρήστες με παρεμφερή ενδιαφέροντα. Ένας άλλος τρόπος αντί της χρήσης ολόκληρων των URLs είναι η χρήση των λεγόμενων *hosts*. Πιο συγκεκριμένα, ο διαφημιστής μπορεί να θέλει η διαφήμιση του να εμφανίζεται όταν ένα host, για παράδειγμα το www.imdb.com, εμφανίζεται στη σελίδα των αποτελεσμάτων αναζήτησης. Το Output URL bidding είναι κάτι διαφορετικό από το keyword bidding, όπου οι διαφημιστές τοποθετούν την προσφορά τους στις λέξεις κλειδιά, που έπειτα οδηγούν σε ιστοσελίδες. Το URL bidding είναι ένας τρόπος για να αναφερόμαστε συνολικά σε όλες αυτές τις λέξεις κλειδιά.

Για παράδειγμα, ας πούμε ότι ο διαφημιστής θέλει να δημιουργήσει μία καμπάνια που να αφορά τους επερχόμενους Ολυμπιακούς Αγώνες. Ο διαφημιστής θα πρέπει να σκεφτεί τα URLs που αποτελούν την πιο αυθεντική και ταιριαστή πηγή πληροφορίας για το γεγονός αυτό. Δηλαδή, όσο πιο καλά μπορέσει να προβλέψει ο διαφημιστής την συμπεριφορά του χρήστη ως προς τα αιτήματα που τυχόν θα υποβάλλει, τόσο πιο επιτυχημένη θα είναι και η καμπάνια του. Στη συνέχεια παρουσιάζεται η διαδικασία για να επιτευχθεί το output URL bidding.

2.1.1 Περιγραφή μοντέλου και ορολογίας

q: query, αυτό που ψάχνει ο χρήστης, δηλαδή ένα αίτημα

U (q): ταξινομημένη λίστα με τα n υψηλότερα URLs που προκύπτουν από την αναζήτηση. Παράδειγμα ενός URL είναι το εξής: en.wikipedia.org/wiki/The_Social_Network

Τυπική τιμή για το n = |U(q)| = 10

α = ad, η διαφήμιση

Για κάθε διαφήμιση α, παρέχεται από τον διαφημιστή μια λογική έκφραση (output expression) που δείχνει πότε μια διαφήμιση μπορεί να είναι υποψήφια προς παρουσίαση.

Παράδειγμα λογικής έκφρασης : $\alpha = u1 \vee u2$, όπου $u1,u2 \in U(q)$

A: ένα σύνολο από τις διαθέσιμες διαφημίσεις

m (a, q): συνάρτηση ταιριάσματος. Αξιολογεί την λογική έκφραση α με το q, δηλαδή ελέγχει αν τα URL του U(q), ικανοποιούν τη λογική έκφραση α.

Συνεπώς, για κάθε q, το σύστημα που υλοποιεί sponsored search, καταλήγει στο: $A(q) = \{ a | a \in A \wedge m(a,q) \}$.

2.1.2 Κατάταξη διαφημίσεων (Ad indexing)

Η μηχανή αναζήτησης χρησιμοποιεί μια λίστα προκειμένου να φέρνει και να τοποθετεί κάθε διαφήμιση α ϵ A από τις διευθύνσεις URL , u ϵ U(q). Για παράδειγμα, υποθέτουμε ότι το σύνολο A περιέχει τρείς διαφημίσεις, δηλαδή A = {α1,α2,α3}. Η μηχανή αναζήτησης χρησιμοποιεί την παρακάτω λίστα I(U->A) ή I(H->A) αν δουλεύουμε με hosts, προκειμένου να ανακτήσουμε όλες τις διαφημίσεις που αντιστοιχούν στις διευθύνσεις URL της λίστας Uq. Έτσι, διαμορφώνεται ο παρακάτω πίνακας:

$$
\begin{aligned}
A &= \{a_1, a_2, a_3\} \\
a_1 &= u_1 \vee u_2 \vee u_3 \\
a_2 &= u_3 \vee u_4 \\
a_3 &= u_2 \vee u_4
\end{aligned}
$$

(a) Ads

URL	Ads
u_1	a_1
u_2	a_1, a_3
u_3	a_1, a_2
u_4	a_2, a_3

(b) Index

Τα κλειδιά για τη δεικτοδότηση είναι τα URLs, και μια λίστα για τα URL ανά κλειδί u περιέχει όλες τις διαφημίσεις. Αν έχουμε να κάνουμε με συνδυασμό διαφημίσεων, η μηχανή αναζήτησης χρησιμοποιεί την ίδια δεικτοδότηση για να βρει για κάθε συνδυασμό την τομή της ένωσης του συνόλου των URLs. Για παράδειγμα, έχουμε (u1 and u2) και έτσι η μηχανή αναζήτησης θα πρέπει να βρει την τομή των λιστών [α1] και [α1,α3] που αντιστοιχούν στο u1 και στο u2 αντίστοιχα.

Στην περίπτωση του output URL bidding, η αρχιτεκτονική που χρησιμοποιούμε σε αντιστοιχία με το γενικότερο σχήμα που παρουσιάσαμε παραπάνω (σχήμα 2) είναι η εξής: η είσοδος στο στοιχείο που αφορά το sponsored search είναι τα URLs και αντιστοιχούν στα αποτελέσματα στη σελίδα των αποτελεσμάτων αναζήτησης.

Όσον αφορά τη διαδικασία του σχήματος 3, πέρα από τα όσα ήδη αναφέρθηκαν, να σημειωθεί ότι η ανάγκη για σειριοποίηση απαιτείται, διότι το στοιχείο που αφορά το sponsored search παρέχει τα URLs που εμφανίζονται σε υψηλή κατάταξη στα αποτελέσματα της αναζήτησης.

2.1.3 Απόκρυψη πληροφορίας (caching)

Για ένα αίτημα q, η μηχανή αναζήτησης πρέπει να ανακτήσει τα URLs που βρίσκονται στο σύνολο U(q), καθώς και το σύνολο διαφημίσεων A(q) που σχετίζονται με το σύνολο U(q). Προκειμένου να αποφύγουμε τους υπολογισμούς για την ανάκτηση των παραπάνω συνόλων, κάθε φορά που ο χρήστης έχει το ίδιο αίτημα, υπολογίζουμε τα σύνολα αυτά μία(1) φορά και έπειτα κρύβουμε (cache) τα αποτελέσματα. Το μόνο που θα χρειάζεται είναι απλά μια αναζήτηση σε έναν κατάλογο - λίστα. Τέλος, χρειάζεται χώρος, αλλά και συχνή ενημέρωση του καταλόγου αποθήκευσης.

Δύο προσεγγίσεις θα μπορούσαμε να παρουσιάσουμε. Στην πρώτη, το στοιχείο sponsored search μπορεί να αποκρύψει έναν κατάλογο αιτημάτων που αφορούν το URL, I(Q->U) και χρησιμοποιεί τον κατάλογο I(U->A). Όταν λαμβάνεται ένα καινούργιο αίτημα, η μηχανή αναζήτησης ανακτά όλες τις διευθύνσεις URL που εμφανίζονται στα αποτελέσματα αναζήτησης, με μια αναζήτηση στον κατάλογο I(Q->U). Στη συνέχεια, χρησιμοποιεί τον

κατάλογο I(U->A) για να βρει τις διαφημίσεις από το σύνολο A(q) που αντιστοιχούν σε κάθε URL. Στη δεύτερη προσέγγιση, το στοιχείο sponsored search αποκρύπτει το I(Q->A). Με χρήση αυτού, η μηχανή αναζήτησης ανακτά διαφημίσεις με μια απλή αναζήτηση. Συνεχίζοντας, χρησιμοποιούμε LRU τεχνική, δηλαδή απορρίπτουμε το λιγότερο πρόσφατα χρησιμοποιούμενο αίτημα από αυτά που έχουμε αποκρύψει.

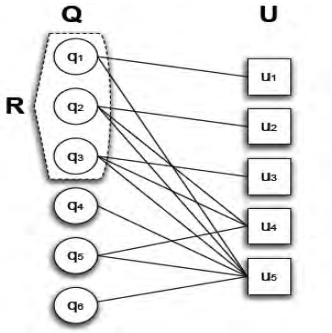
Σε τελευταίο στάδιο, αναφερόμαστε στο πώς διαχειριζόμαστε μια περίπτωση αποτυχίας απόκρυψης (cache miss handling). Σε αυτήν την περίπτωση υπάρχουν δύο εναλλακτικές. Στην πρώτη, η μηχανή αναζήτησης μπορεί να αγνοήσει το cache miss και έτσι να μην εμφανιστούν διαφημίσεις στον χρήστη που είχε αυτό το αίτημα q. Στην δεύτερη περίπτωση, η μηχανή αναζήτησης μπορεί να ανακτήσει τις διαφημίσεις από το σύνολο A(q), χρησιμοποιώντας μια από τις μεθόδους που περιγράφηκαν παραπάνω και αφορούν το σχήμα 3.

2.1.4 Εξυπηρέτηση αιτημάτων μέσω έκφρασης (Query set output cover)

Έχουμε ένα σύνολο R που περιέχει τα αιτήματα που ενδιαφέρουν έναν διαφημιστή. Θέλουμε να αναπαραστήσουμε την πρόθεση των διαφημιστών με μια έκφραση εξόδου α, η οποία να περιέχει λιγότερα URLs από τα αιτήματα που υπάρχουν στο σύνολο R. Λέμε ότι η έκφραση α καλύπτει (cover) το σύνολο αιτημάτων R, αν για κάθε $q \epsilon R$, η συνάρτηση ταιριάσματος m(a,q) είναι αληθής. Επίσης, θέλουμε η έκφραση α να είναι συμπαγής (compact), δηλαδή να είναι ικανή να περιγράψει τις σελίδες που ενδιαφέρουν τους διαφημιστές με όσο το δυνατόν λιγότερα URLs. Για να ορίσουμε αυτή την ιδιότητα των εκφράσεων, ορίζουμε την εξής σχέση: |α| = το πλήθος των διαζεύξεων στο σύνολο α. Για παράδειγμα, το μέγεθος της έκφρασης (u1∨u2) είναι δύο(2). Τέλος, ένα ακόμη χαρακτηριστικό είναι ότι επιδιώκεται το α να αντιστοιχίζει λιγότερα αιτήματα από αυτά που υπάρχουν στο σύνολο R.

2.1.5 Η ιδιότητα Spill

Μια έκφραση εξόδου η οποία καλύπτει το σύνολο R, μπορεί να αντιστοιχίζει αιτήματα που δεν ανήκουν στο R. Αναφερόμαστε στο σύνολο αυτών των αιτημάτων ως: spill(a,R) = {q : q ∈ (Q - R) ∧ m(a,q)}. Το σύνολο Q περιέχει όλα τα αιτήματα που χρειαζόμαστε για τον υπολογισμό του spill και μπορούμε να τα αποκτήσουμε από τις καταγραφές (logs) των μηχανών αναζήτησης. Προκειμένου να υπολογίσουμε το spill μιας έκφρασης α, αναπαριστούμε τα αιτήματα του συνόλου Q και τα αποτελέσματα της αναζήτησης με το παρακάτω γράφημα :



Σχήμα 4 : Γράφημα υπολογισμού της ιδιότητας Spill [Πηγή: από Βιβλ. Κεφ., αναφορά 1].

Οι κόμβοι του γραφήματος U αναπαριστούν τις διευθύνσεις URL και το Q αναπαριστά τα αιτήματα των χρηστών μαζί με τα αποτελέσματα της μηχανής αναζήτησης. Η ακμή (q1,u1) δείχνει ότι το URL u1 εμφανίζεται στα υψηλότερα σε βαθμίδα αποτελέσματα που αφορούν το αίτημα q1. Στη συνέχεια, στο σύνολο R = {q1,q2,q3}, κάποιες από τις εκφράσεις που καλύπτουν το σύνολο R και τα αντίστοιχα σύνολα spill είναι τα παρακάτω:

$$a_1 := u_1 \vee u_2 \vee u_3 \qquad spill(a_1, R) = \{\}$$
$$a_2 := u_1 \vee u_4 \qquad spill(a_2, R) = \{q_5\}$$
$$a_3 := u_5 \qquad spill(a_3, R) = \{q_4, q_5, q_6\}$$

Το spill για την έκφραση α2 είναι το {q5}, επειδή το α2 αντιστοιχίζεται με τα αιτήματα q1,q2,q3,q5 και τα πρώτα τρία αιτήματα βρίσκονται στο σύνολο R. Γενικά, επιδιώκουμε την ελαχιστοποίηση του spill.

2.1.6 Πρόβλημα βελτιστοποίησης

Αυτό που θέλουμε είναι να ελαχιστοποιήσουμε το εξής: minimize γ|α| + (1-γ)|spill(a,R)|, έχοντας τη συνάρτηση ταιριάσματος m(a, q) για κάθε q ∈ R. Το γ ∈ [0,1] είναι παράμετρος κανονικοποίησης ανάμεσα στις δύο ιδιότητες που έχουμε αναφέρει παραπάνω, compactness και spill. Για παράδειγμα, αν το γ=1, η έκφραση είναι πιο συμπαγής κοκ. Το πρόβλημα του query set output cover είναι αρκετά δύσκολο να επιλυθεί. Πιο συγκεκριμένα, είναι *NP-hard* να βρεθεί η πιο συμπαγής (compact) έκφραση που να καλύπτει όλα τα αιτήματα του συνόλου R, αλλά και να καταλήγει σε όσο το δυνατόν μικρότερο spill.

2.1.7 Ο αλγόριθμος για το πρόβλημα query set output cover

Algorithm 1 Greedy Query Set Output Cover

Input: R, γ, Q
Output: a, $spill$

Allocate empty hash tables C and S. $C[u]$ (or $S[u]$) returns $\emptyset$ if key u is not in C (or S).
for all $q \in Q$ **do**
  for all $u \in U(q)$ **do**
    if $q \in R$ **then**
      $C[u] \leftarrow C[u] \cup \{q\}$
    else
      $S[u] \leftarrow S[u] \cup \{q\}$
$R^{rem} \leftarrow R$ ▷ R^{rem}: Remaining queries to be covered
$U_a \leftarrow \emptyset$ ▷ U_a: URLs in logical expression a
$spill \leftarrow \emptyset$
while $R^{rem} \neq \emptyset$ **do**
  $u \leftarrow arg\max_{u' \notin U_a} |C[u'] \cap R^{rem}| - w(\gamma)|S[u'] - spill|$
  $U_a \leftarrow U_a \cup \{u\}$
  $R^{rem} \leftarrow R^{rem} - C[u]$
  $spill \leftarrow spill \cup S[u]$
$a \leftarrow$ DISJUNCTION(U_a)
return $a, spill$

Περιγράφεται η greedy (άπληστη) προσέγγιση για να βρεθεί μια έκφραση α που να είναι διάζευξη διευθύνσεων URL. Είσοδο του αλγόριθμου αποτελούν το σύνολο αιτημάτων R, η

τιμή της παραμέτρου γ και το σύνολο αναφοράς Q. Ο αλγόριθμος επιστρέφει την έκφραση εξόδου α και το αντίστοιχο spill που αφορά το σύνολο R. Στις γραμμές 1-7, ο αλγόριθμος υπολογίζει για κάθε URL u που εμφανίζεται στα αποτελέσματα κάποιου αιτήματος q ∈ R, τα ακόλουθα σύνολα:

$C[u] = \{q |\ q \in R \wedge u \in U(q)\}$ R αιτήματα καλύπτονται από το u

$S[u] = \{q |\ q \in (Q-R) \wedge u \in U(q)\}$ το spill του URL u

Σε κάθε βήμα προσθέτουμε ένα μοναδικό URL.

Στην κυρίως επανάληψη, στις γραμμές 11-15, ο αλγόριθμος υπολογίζει τις διαζεύξεις με τρόπο που να αυξάνεται σε κάθε βήμα το σύνολο Uα της έκφρασης εξόδου α, καθώς επίσης και το spill. Σε κάθε επανάληψη το URL u που προστίθεται στο σύνολο Uα επιλέγεται με greedy (άπληστο) τρόπο, λαμβάνοντας υπόψη τα εξής:

α) $|C[u] \cap R^{rem}|$: ο αριθμός των αιτημάτων που δεν έχουν καλυφθεί όταν συμβαίνει το u

β) $|S[u] - spill|$: ο αριθμός των επιπρόσθετων spill αιτημάτων που προκαλεί το u

Το URL u θα πρέπει να μεγιστοποιεί το α και να ελαχιστοποιεί το β. Έτσι, ο αλγόριθμος επιλέγει τέτοιο u ώστε να μεγιστοποιεί τη διαφορά ανάμεσα στους δύο αριθμούς, λαμβάνοντας υπόψη το επιπλέον spill με βάρος w(γ). Η τιμή του βάρους w(γ) θα πρέπει να αντικατοπτρίζει σε αυτήν την τοπική απόφαση την επιθυμητή διαφορά ανάμεσα στις δύο ιδιότητες, compactness και spill. Ορίζουμε το w(γ) ως: w(γ) = (1-γ)/γ προκειμένου να ικανοποιεί τις εξής τρεις ιδιότητες α) η w(γ) να είναι μια φθίνουσα συνάρτηση του γ β) η τιμή w(0) = ∞ και γ) η τιμή w(1) = 0. Στο τέλος κάθε επανάληψης, ο αλγόριθμος ανανεώνει την τιμή του spill και τα αιτήματα R^{rem} , που είναι τα υπολειπόμενα αιτήματα που πρέπει να καλυφθούν από τις επόμενες επαναλήψεις. Τέλος, στις γραμμές 16-17, ο αλγόριθμος υπολογίζει το α που είναι η διάζευξη όλων των URL στο Uα και τελικά επιστρέφει το α και την τιμή του spill.

Όσον αφορά κάποια επιπλέον στοιχεία για την υλοποίηση του αλγόριθμου, αυτός χρησιμοποιεί τα εξής σύνολα $|C[u'] \cap R^{rem}|$ και $|S[u'] - spill|$, για να επιλέξει το URL που θα προστεθεί στο Uα. Η πραγματική υλοποίηση όμως, με κάποια γλώσσα προγραμματισμού θα χρησιμοποιήσει αντί των συνόλων, μετρητές που θα ανανεώνονται σε κάθε επανάληψη. Στη γραμμή 12 του αλγόριθμου, που είναι το βήμα της επιλογής, θα χρειασθεί στη χειρότερη περίπτωση $\log|U|$ χρόνος. Αν ο βαθμός του επιλεγμένου URL u είναι d(u), τότε ο αλγόριθμος θα πρέπει να ανανεώσει τους μετρητές όλων των URLs στα αποτελέσματα αυτών των d(u) αιτημάτων. Αυτό θα χρειασθεί $nd(u)\log|U|$ χρόνο στη χειρότερη περίπτωση.

Ο αλγόριθμος μπορεί να χρησιμοποιηθεί και για hosts εκτός από URLs, αλλά και ταυτόχρονα για URL και hosts υπολογίζοντας τα σύνολα C[u] και S[u] για την ένωση των URLs και των hosts.

2.1.8 Πειράματα – Παραδείγματα

Σε τελευταίο στάδιο θα παρουσιαστεί μια προσπάθεια αξιολόγησης της διαδικασίας του Output URL bidding η οποία βασίζεται σε εκφράσεις εξόδου που παράγονται με χρήση του αλγόριθμου που μελετήσαμε. Σκοπός είναι να ελεγχθεί η ιδιότητα compactness και spill, αλλά και γενικότερα η λειτουργικότητα του αλγόριθμου για το πρόβλημα query set output cover. Τα δεδομένα μας προέρχονται από μιας μέρας αιτήματα χρηστών στη μηχανή αναζήτησης της Yahoo τη χρονική περίοδο του Νοέμβρη 2009. Για κάθε αίτημα χρήστη τα αρχεία περιέχουν τα URLs των μεγαλύτερων δέκα(10) αποτελεσμάτων, αλλά επίσης και όλες τις διαφημίσεις που είδαν οι χρήστες και υπέβαλαν κάποιο αίτημα. Για την παραγωγή μιας έκφρασης εξόδου ακολουθείται η εξής διαδικασία: Πρώτα, διαλέγουμε μια διαφήμιση από αυτές που έχουμε στο υλικό του πειράματος και ορίζουμε το R σαν το σύνολο με τα αρχεία αιτημάτων όπου η διαφήμιση εμφανίστηκε. Στη συνέχεια, παρέχουμε το σύνολο R σαν είσοδο στον αλγόριθμο προκειμένου να πάρουμε σαν αποτέλεσμα μια φράση εξόδου που να καλύπτει όλα τα αιτήματα του συνόλου R.

Η υλοποίηση του αλγόριθμου έγινε σε γλώσσα προγραμματισμού Python και χρησιμοποιήθηκε το PyLucene για να δεικτοδοτηθούν τα αιτήματα κάθε διαφήμισης, τα URLs και το host κάθε αιτήματος. Παρατηρήθηκε ότι η τιμή του γ επηρέασε τον χρόνο τρεξίματος του αλγόριθμου, και όταν η τιμή του ήταν μεγάλη, είχε ως αποτέλεσμα μεγάλο spill το οποίο αυξανόταν σε κάθε επανάληψη του αλγόριθμου.

Όσον αφορά μια έκφραση εξόδου λέμε ότι καλύπτει ένα συγκεκριμένο σύνολο R και μπορεί να έχει είτε μικρό, είτε μεγάλο spill, ανάλογα με την τιμή του γ που θα χρησιμοποιηθεί. Ένα από τα αποτελέσματα που προέκυψε από τα πειράματα που έγιναν έδειξε ότι οι εκφράσεις εξόδου που βασίζονται στα URLs και στα hosts μαζί, είναι περισσότερο συμπαγείς (compact) και προκαλούν μικρότερο spill, συγκριτικά με εκφράσεις που βασίζονται μόνο σε URLs ή μόνο σε hosts. Για παράδειγμα, εκφράσεις εξόδου μπορεί να είναι περισσότερο συμπαγείς από εκφράσεις που σχετίζονται με URLs, γιατί ένα host αντιστοιχίζει όλα τα αποτελέσματα αναζήτησης με τα URLs από αυτή τη σελίδα host. Όμως, το να αντιστοιχιστούν πολλά αιτήματα χρήστη έχει αρνητική επιρροή στην ιδιότητα spill. Για μικρές τιμές του γ της τάξης του 0.9999, οι εκφράσεις εξόδου περιέχουν πολύ δημοφιλή hosts όπως για παράδειγμα το en.wikipedia.org, το οποίο αντιστοιχίζει χιλιάδες αιτήματα και έτσι το μέσο spill είναι αρκετά μεγάλο. Για ακόμα μικρότερες τιμές του γ, το μέγεθος του spill είναι ακόμη υψηλό, γιατί πολλά αιτήματα περιέχουν μόνο δημοφιλή hosts και έτσι δεν μπορούν να μην συμπεριληφθούν στη λογική έκφραση.

Για τις εκφράσεις εξόδου που βασίζονται σε URLs παρατηρείται ότι πολλά URLs εμφανίζονται μόνο μια φορά σε όλα τα αποτελέσματα αναζήτησης και πάνω από τα μισά αιτήματα του συνόλου R περιέχουν τέτοια URLs. Ως αποτέλεσμα θα πρέπει οι εκφράσεις URL να περιέχουν συνδέσεις που να καλύπτουν μόνο ένα αίτημα και αυτό επηρεάζει φυσικά και την ιδιότητα compactness. Τέλος, αν χρησιμοποιηθεί ένας συνδυασμός εκφράσεων, θα συνδυαστούν τα πλεονεκτήματα και των hosts και των URLs. Μπορεί να επιτευχθεί μικρό μέγεθος με τη χρήση διάσημων hosts, αλλά μπορεί να ελαχιστοποιείται και το spill με τη χρήση πιο σπάνιων URLs.

Ένα δεύτερο αποτέλεσμα από τα πειράματα έδειξε ότι οι εκφράσεις εξόδου που χρησιμοποιούν περισσότερες από δύο συνδέσεις π.χ. k=2 προκαλούν λιγότερα αιτήματα spill από εκφράσεις που δεν έχουν πολλές συνδέσεις, π.χ. k=1. Για παράδειγμα, τα επικρατέστερα hosts για στίχους τραγουδιών σε αποτελέσματα αιτημάτων είναι περίπου τα ίδια π.χ. www.metrolyrics.com, www.last.fm, κλπ. Συνεπώς, χωρίς να έχει σημασία πόσες και ποιες συνδέσεις θα επιλέξουμε για να καλύψουμε ένα αίτημα χρήστη που αφορά στίχους τραγουδιών, η σύνδεση που θα δημιουργηθεί θα ταιριάξει και άλλα αιτήματα χρήστη με το ζητούμενο των στίχων. Έτσι, ο αριθμός των αιτημάτων spill, δεν μειώνεται με χρήση εκφράσεων που βασίζονται σε hosts.

Ένα τρίτο αποτέλεσμα δείχνει συνοπτικά ότι όταν το μέγεθος του αρχικού συνόλου |R| των αιτημάτων αυξάνεται, η έκφραση εξόδου που το καλύπτει είναι πιο συμπαγής. Αναλυτική παρουσίαση και ανάλυση των αποτελεσμάτων στη σχετική πηγή. [Πηγή: από Βιβλ. Κεφ., αναφορά 1]

2.2 Αυτόματη παραγωγή φράσεων προσφοράς (Automatic generation of bid phrases)

Στην ενότητα αυτή, θα παρουσιαστούν κάποιοι μέθοδοι προκειμένου να παραχθούν φράσεις προσφοράς (bid phrases) απευθείας από τις σελίδες, δηλαδή εκφράσεις που θα αποτελέσουν την προσφορά του διαφημιστή. Για να παραχθεί μια διαφήμιση κειμένου, ο διαφημιστής πρέπει να φτιάξει το κείμενο της διαφήμισης και να το συνδέσει με μια σελίδα προορισμού (landing page) που θα περιγράφει το προϊόν ή την υπηρεσία που προωθείται. Ακόμη, ο διαφημιστής θα πρέπει να αντιστοιχίσει τη διαφήμιση του σε ένα σύνολο από επιλεγμένες εκφράσεις (bid phrases), οι οποίες θα αναπαριστούν τα αιτήματα του χρήστη και θα ενεργοποιούν στη συνέχεια τη διαφήμιση. Η κύρια προσέγγιση χρησιμοποιεί ένα παραγωγικό πρότυπο με χρήση ενός πλαισίου μετάφρασης, ώστε το σύστημα να μεταφράζει οποιαδήποτε σελίδα (landing page) που περιέχει τις διαφημίσεις του διαφημιστή, σε σχετικές φράσεις, τις λεγόμενες bid phrases. Αυτές οι φράσεις που παράγονται δεν θα πρέπει να περιέχονται στην είσοδο που δίνει ο χρήστης. Το μοντέλο λοιπόν αυτό, χρησιμοποιεί δύο βασικά στοιχεία. Ένα στοιχείο που έχει να κάνει με τη γλώσσα, για να επιλέγει καλά δομημένες εκφράσεις, και το άλλο στοιχείο αφορά τη μετάφραση, και βοηθά να παραχθούν νέες εκφράσεις που μπορεί και να μην εμφανισθούν στη σελίδα. Είναι επιθυμητό αυτή η έκφραση να είναι όσο το δυνατόν μεγαλύτερη, αλλά από την άλλη πλευρά είναι επίσης καλό αυτές οι φράσεις να είναι σχετικές με το προϊόν ή την υπηρεσία, γιατί αλλιώς οι χρήστες δεν θα ανταποκριθούν.

Σε δεύτερη φάση, οι υποψήφιες εκφράσεις κατατάσσονται σε ένα πιθανοτικό πλαίσιο, χρησιμοποιώντας τα δύο μοντέλα που περιγράψαμε παραπάνω.

Εμπειρική χρήση βασισμένη στην πραγματικότητα, στην κατασκευή σελίδων προορισμού από διαφημιστές και στη διασύνδεσή τους με λέξεις προσφοράς, επιβεβαιώνει την αξία της προσέγγισής μας. Με επιτυχία έχουμε την παραγωγή πολλών φράσεων προσφοράς από τις μεθόδους που περιγράψαμε, που φυσικά δίνουν καλύτερα αποτελέσματα, συγκρινόμενη με την μέθοδο εξόρυξης φράσεων προσφοράς απλά από το κείμενο μιας διαφήμισης.

2.2.1 Μέθοδος

l: μια ιστοσελίδα που μπορεί να χρησιμοποιηθεί σαν σελίδα προορισμού (landing page) μιας διαφήμισης

b: bid phrase, η φράση προσφοράς

Σκοπός είναι να παραχθούν αυτόματα μία ή περισσότερες εκφράσεις προσφοράς b για την σελίδα προορισμού (landing page) l. Έχουμε τα σύνολα λέξεων b = {b1,b2,…,bn} και l = {l1,l2,…,lm} και όπου τα bi, lj υποδηλώνουν λέξεις. Να σημειωθεί ότι η φράση b μπορεί και να μην υπάρχει σε μια σελίδα προορισμού l, δηλαδή μπορεί όλες οι λέξεις στο σύνολο b να μην προέρχονται από το l.

2.2.2 Κατάταξη υποψήφιων εκφράσεων (ranking candidate phrases)

Pr(l/b): η πιθανότητα να παράγω l ιστοσελίδες από το σύνολο b

Θέλουμε να κατατάξουμε τις εκφράσεις στηριζόμενοι στην πιθανότητα Pr(b/l), δηλαδή η πιθανότητα να εμφανισθεί μια φράση b δεδομένης μιας landing page l. Εφαρμόζουμε τον νόμο του Bayes και έτσι έχουμε : $\Pr(b/l) = \frac{\Pr(l/b)\Pr(b)}{\Pr(l)}$.

Δηλαδή, έχουμε δύο ανεξάρτητα μεταξύ τους στοιχεία τα Pr(l/b) και Pr(b). Η πιθανότητα Pr(l/b) ονομάζεται μοντέλο μετάφρασης (translation model,TM), αφού είναι οι πιθανότητες μετάφρασης από μια bidding phrase σε μια landing page. Η πιθανότητα Pr(b), ονομάζεται μοντέλο γλώσσας της έκφρασης (bid phrase language model, LM), αφού χαρακτηρίζει πότε μια φράση είναι πιθανόν να είναι μια έγκυρη φράση προσφοράς.

2.2.3 Το μοντέλο μετάφρασης (translation model, TM)

Σκοπός είναι να γεφυρωθεί η απόσταση που έχει να κάνει με το λεξιλόγιο, ώστε να δίνεται κάποιος θετικός βαθμός για φράσεις που είναι σχετικές με την σελίδα.

Θα γίνει αρχικά η κατασκευή συνόλων που θα είναι μια συλλογή από σελίδες προορισμού (landing pages), σε συνδυασμό με τις εκφράσεις προσφορών που παρέχουν οι διαφημιστές. Έτσι, για κάθε σελίδα l και για κάθε φράση b που συνδέεται με αυτήν, παράγουμε ένα στιγμιότυπο: b1 b2 … bn -> l1 l2 … lm.

Για να γίνει κατανοητή η παραπάνω διαδικασία, θα δοθεί ένα παράδειγμα. Για μια σελίδα προορισμού η οποία πραγματεύεται θέματα τα οποία σχετίζονται με τον ηθοποιό Kirsten Dunst, πιθανόν να περιέχει λέξεις όπως: Kirsten, dunst, film, gossip, girl,… και οι φράσεις προσφοράς που προέρχονται από τον διαφημιστή πιθανόν να είναι οι εξής: «Kristen dunst story»,…κοκ. Κάθε μία από αυτές τις φράσεις προσφοράς, ζευγαρώνει με τις λέξεις της σελίδας προορισμού προκειμένου να δημιουργήσει ένα παράλληλο στιγμιότυπο δοκιμής για το μοντέλο μετάφρασης (translation model) όπως ακολουθεί: Kristen dunst movie -> Kristen dunst film gossip …, Kristen dunst interview -> Kristen dunst film gossip …,Kristen dunst story-> Kristen dunst film gossip … Κάθε φράση δηλαδή, ζευγαρώνει με τις λέξεις της σελίδας προορισμού και έτσι φτιάχνει ένα καινούργιο στιγμιότυπο του μοντέλου μετάφρασης. Για να μικρύνει το κενό (gap) του λεξιλογίου ανάμεσα σε φράσεις και σελίδες

προορισμού, χρησιμοποιείται πολλές φορές το περιεχόμενο της διαφήμισης που αντιστοιχεί με τη συγκεκριμένη σελίδα (bid phrase – ad word content) και προκύπτουν καινούργια ζευγάρια ως ένα επιπλέον στιγμιότυπο.

Επίσης, ουσιαστικά αυτό που θέλουμε είναι να υπολογίσουμε το μοντέλο μετάφρασης, δηλαδή να ορίσουμε την πιθανότητα Pr(l/b). Έτσι, έχουμε: $\Pr(l|b) \propto \prod_j \sum_i t(l_j|b_i)$. Όπου t είναι ο πίνακας μετάφρασης, ο οποίος χαρακτηρίζει την πιθανότητα ένα στοιχείο της σελίδας προορισμού να έχει παραχθεί από ένα στοιχείο της φράσης προσφοράς (bid phrase). Από τη στιγμή που έχουμε έναν πίνακα γεμάτο από λέξεις τόσο για το λεξιλόγιο των φράσεων, όσο και για το λεξιλόγιο των σελίδων, μπορούμε να υπολογίσουμε την πιθανότητα Pr(l/b) για τα ζεύγη που δεν έχουμε συναντήσει ακόμη. Για να υπολογίσουμε το t (lj/bi) , θα πρέπει να υποθέσουμε ότι κάθε λέξη του συνόλου b έχει ίδια πιθανότητα να ταιριάζει με όλες τις λέξεις της σελίδας προορισμού. (Παραπομπή στο IBM model 1 από Πηγή: Βιβλ. Κεφ., αναφορά 2).

Για παράδειγμα, σε συνέχεια του παραδείγματος με τον ηθοποιό που αναφέραμε πιο πάνω, υποθέτουμε ότι γνωρίζουμε ότι η λέξη film στη σελίδα προορισμού έχει παραχθεί από τη λέξη movie της φράσης προσφοράς (bid phrase) ή ακολουθώντας τις λέξεις κατά τη διαδικασία μετάφρασης, ότι η λέξη film θα πρέπει να ευθυγραμμιστεί με τη λέξη movie. Αν μας δινόταν αυτή η πληροφορία εξίσωσης της λέξης για όλες τις εμφανίσεις της λέξεις movie, ο υπολογισμός του t(film|movie) είναι κάτι τετριμμένο. Ουσιαστικά, μετράμε τον αριθμό των φορών που η λέξη movie ευθυγραμμίζεται με τη λέξη film, καθώς επίσης και τον αριθμό των φορών που η λέξη movie ευθυγραμμίζεται με οποιαδήποτε άλλη λέξη στις σελίδες προορισμού, και τελικά υπολογίζουμε το κλάσμα μεταξύ αυτών των δύο υπολογισμών.

Χρειάζεται να σημειωθούν κάποια πράγματα που αφορούν τα άδεια στοιχεία (null tokens). Αυτά χρησιμοποιούνται ώστε να λαμβάνουμε υπόψη μας λέξεις που δεν ταιριάζουν καλά με κάποια λέξη από την άλλη πλευρά. Στην περίπτωσή μας, χρησιμοποιούνται για εκείνα τα lj που δεν σχετίζονται στενά με το σύνολο b. Ένα παράδειγμα το οποίο θα βοηθήσει την κατανόηση της παραπάνω διαδικασίας είναι αυτό που ακολουθεί. Θα αναλογιστούμε ένα σχήμα της μορφής: Honda -> best car dealer, mp3 player -> buy a new iPod, Honda -> Buy a new car, mp3 player -> best price on nano. Προκύπτει από το μοντέλο μετάφρασης ο ακόλουθος πίνακας:

b_i	translations			
	l_j	$t(l_j \mid b_i)$	l_j	$t(l_j \mid b_i)$
mp3	*on*	0.18	*price*	0.18
	nano	0.18	*a*	0.07
	ipod	0.18	*Best*	0.07
	price	0.18	*new*	0.07
Honda	*car*	0.49	*a*	0.07
	dealer	0.25	*Best*	0.07
	a	0.07	*new*	0.07
null	*a*	0.24	*car*	0.03
	Best	0.24	*dealer*	0.01
	new	0.24	*nano*	0.00
	Buy	0.24	*ipod*	0.00
	car	0.03	*on*	0.00

Σημειώνεται ότι οι περισσότερες λέξεις που δεν δίνουν κάποια πληροφορία στις σελίδες προορισμού συνήθως λαμβάνονται υπόψη σαν κενές λέξεις (null tokens). Έτσι, ο πίνακας μετάφρασης θα μπορεί να εστιάσει σε περισσότερο σημαντικές λέξεις όπως car ή iPod. Για τον υπολογισμό της πιθανότητας Pr(li|bj) = count(lj,bi)/count(bi), έχουμε ότι Pr(best|Honda) = 1/2Pr(car|Honda). Με την εισαγωγή των κενών λέξεων (null tokens) παράγεται ένας πιο λογικός πίνακας μετάφρασης.

Τέλος, κάποιες λέξεις στη σελίδα (landing page) είναι πιο σημαντικές από άλλες. Για παράδειγμα, έχοντας μια σελίδα HTML, λέξεις που εμφανίζονται στον τίτλο κλπ είναι χαρακτηριστικές για τη σελίδα. Έτσι, δίνουμε ένα βάρος wj για κάθε lj ϵ l. Μπορούμε λοιπόν, να δώσουμε μικρό βάρος σε κανονικού περιεχομένου λέξεις και μεγαλύτερο βάρος σε λέξεις με σημαντικότερο περιεχόμενο. Επομένως, ο υπολογισμός της πιθανότητας τώρα γίνεται: $\Pr(\mathrm{l/b}) \propto \prod_j (\sum_i \mathrm{t(lj/bi)})^{wj}$. Στη συνέχεια, μειώνουμε το σύνολο lj στα n μεγαλύτερα ως προς το βάρος wj στοιχεία. Με αυτόν τον τρόπο το ταίριασμα γίνεται με τις πιο σημαντικές λέξεις τις σελίδας.

2.2.4 Το μοντέλο γλώσσας (bid phrase language model)

Οι διαφημιστές επιλέγουν συχνά φράσεις που θα τις χρησιμοποιήσουν για την προσφορά τους, οι οποίες ταιριάζουν με δημοφιλή αιτήματα χρηστών. Κάτι τέτοιο γίνεται για να αυξηθούν οι πιθανότητες οι διαφημίσεις τους να εμφανισθούν και να κλικαριστούν από τους χρήστες. Τα αποθηκευμένο υλικό που μπορεί να αντληθεί από τα αρχεία (logs) των μηχανών αναζήτησης είναι μια καλή πηγή καλά δομημένων φράσεων. Πιο συγκεκριμένα, η πιθανότητα Pr(b) = $\prod_i \Pr(b_i/b_{i-1})$, όπου $\Pr(b_i/b_{i-1})$ = λ1f(bi) + λ2f(b_i/b_{i-1}) και με λ1+λ2=1. Ας θεωρήσουμε το c(bi,bj) τον αριθμό των φορών bibj που εμφανίζεται σε μια ακολουθία Q. Έχουμε λοιπόν ότι, f(b_i/b_{i-1}) $= \frac{\mathrm{c(bi-1,bi)}}{\sum_\mathrm{j} \mathrm{c(bi-1,bj)}}$. Το c(i) είναι ο αριθμός των φορών που το bi εμφανίζεται στην ακολουθία Q, δηλαδή είναι ο αριθμός των ανεξάρτητων στοιχείων στο Q. Έτσι, το f(bi) ορίζεται ως εξής: f(b_i) $= \frac{\mathrm{c(bi)+1}}{\sum_\mathrm{j} \mathrm{c(bj)+|V|}}$, το |V| είναι το μέγεθος των λέξεων του Q. Τέλος, το μοντέλο αυτό επιλέγει φράσεις που είναι πιθανό να εμφανισθούν σε ακολουθίες προτάσεων, αλλά και φράσεις που περιέχουν ζευγάρια στοιχείων και είναι πιθανό να εμφανισθούν στα αρχεία αιτημάτων.

2.2.5 Παραγωγή υποψήφιων φράσεων

Γενικά υποψήφιες εκφράσεις θα μπορούσαν να είναι όλες οι δυνατές φράσεις που υπάρχουν σε ένα αρχείο αιτημάτων. Για να κατατάξουμε όλες τις φράσεις b ∈ Q, σύμφωνα με το $\Pr(\mathrm{l/b})\Pr(\mathrm{b})$, είναι κάτι όχι και τόσο εφικτό. Έτσι, ελέγχουμε φράσεις από την σελίδα προορισμού (landing page). Πιο συγκεκριμένα, όταν χρησιμοποιούμε τις landing pages άμεσα, δηλαδή επιλέγοντας λέξεις ή φράσεις που υπάρχουν στη σελίδα ή έμμεσα, δηλαδή χρησιμοποιώντας το μοντέλο μετάφρασης που περιγράψαμε παραπάνω, προκειμένου να παράγουμε καινούργιες υποψήφιες εκφράσεις που θα απαρτίσουν το σύνολο B_{LP}. Όμως, οι διαφημιστές συχνά θέλουν να επιλέγουν φράσεις προσφορών που δεν ανήκουν ως φράσεις στις σελίδες προορισμού. Αρχικά, επιλέγουμε τις n_p λέξεις από τη σελίδα με το μεγαλύτερο βάρος wj. Για κάθε λέξη επιλέγουμε τις nt πιο πιθανές μεταφράσεις από τον πίνακα μεταφράσεων. Αυτές οι λέξεις έπειτα συνδυάζονται σε όλους τους πιθανούς συνδυασμούς

για να αποτελέσουν τις υποψήφιες εκφράσεις. Επίσης, παράγονται και μικρότερες ακολουθίες από αυτές τις εκφράσεις οι οποίες προστίθενται στο σύνολο.

2.2.6 Σύστημα ταιριάσματος περιεχομένου (content match system)

Σκοπός αυτού του συστήματος είναι να βρει τις πιο σχετικές διαφημίσεις δοσμένης μιας σελίδας προορισμού για την παραγωγή υποψήφιων εκφράσεων. Αυτό επιτυγχάνεται μετατρέποντας τη σελίδα σε ένα διάνυσμα με κάποια χαρακτηριστικά, όπου μια ιδιότητα μπορεί να αντιστοιχεί σε μια λέξη ή φράση της σελίδας προορισμού και το βάρος κάθε χαρακτηριστικού να δείχνει τη σημασία κάθε μιας. Αυτή η διαδικασία γίνεται και για τις διαφημίσεις. Στη συνέχεια, το αν σχετίζεται μια διαφήμιση με μια σελίδα προορισμού μπορεί να μετρηθεί εφαρμόζοντας μια συνάρτηση ομοιότητας στα αντίστοιχα χαρακτηριστικά του διανύσματος.

Περιγράφεται στη συνέχεια το πώς μπορούμε να παράγουμε εκφράσεις προσφοράς (bid phrases) για μια νέα σελίδα προορισμού. Αρχικά, εφαρμόζουμε το παραπάνω σύστημα (content match system-CMS) στη σελίδα για να πάρουμε κάποιες υψηλές σε βαθμολογία διαφημίσεις. Το σύστημα κατατάσσει τις διαφημίσεις με βάση την αξία προσφοράς (bid value) και με βάση το πόσο σχετίζεται με τη σελίδα προορισμού. Για κάθε διαφήμιση, διαλέγει τις πιο κατάλληλες και τις μαζεύει σε ένα σύνολο, το Bcms. Στη συνέχεια, κατατάσσει τις φράσεις που βρίσκονται στο σύνολο αυτό. Η κατάταξη μπορεί να γίνει με δύο τρόπους. Ο πρώτος, κατατάσσει τις εκφράσεις με τη σειρά που εμφανίζονται και έχουν καταταχθεί από το σύστημα CMS. Ο δεύτερος τρόπος, τις κατατάσσει με βάση τον αριθμό των εμφανίσεων.

2.2.7 Σύστημα εξόρυξης (Extraction-based system)

Σε πρώτη φάση παράγονται οι υποψήφιες φράσεις και σε δεύτερη φάση γίνεται η κατάταξη τους. Πιο συγκεκριμένα, πρώτα επεξεργαζόμαστε τη σελίδα και το URL προορισμού και βρίσκουμε λέξεις και φράσεις από τη σελίδα. Προσθέτουμε αυτές τις φράσεις στο σύνολο των υποψήφιων φράσεων. Υπολογίζουμε ένα βάρος $w_k = \frac{fk}{\log(Nd)}$ για κάθε λέξη b_k από τις υποψήφιες φράσεις, όπου το fk αναπαριστά τη συχνότητα εμφάνισης της λέξης ανάμεσα στις υποψήφιες και το Nd είναι ο αριθμός των εγγράφων στον παγκόσμιο ιστό (web) που περιέχουν αυτή τη λέξη. Αναπαρίσταται κάθε υποψήφια λέξη σαν ένα διάνυσμα με βάρη. Επίσης, κάποιες «κακές» υποψήφιες φράσεις φιλτράρονται με χρήση ενός ορίου στο βάρος των λέξεων που μπορούν να εμφανισθούν σε μια υποψήφια φράση. Με τον ίδιο τρόπο αναπαρίσταται και η σελίδα προορισμού. Τέλος, υπολογίζουμε την ομοιότατα των δύο διανυσμάτων. Το αποτέλεσμα αυτό χρησιμοποιείται για την κατάταξη του συνόλου των φράσεων προσφοράς για δεδομένη σελίδα προορισμού.

2.2.8 Σύστημα διακριτοποίησης (discriminative system)

Το σύστημα αυτό εκτελεί μόνο κατάταξη των φράσεων προσφοράς. Οι υποψήφιες φράσεις παρέχονται στο σύστημα αυτό με χρήση ενός από τα συστήματα που περιγράφτηκαν συνοπτικά σε προηγούμενες παραγράφους. Το σύστημα διακριτοποίησης υπολογίζει ποικίλα χαρακτηριστικά προκειμένου να κάνει την κατάταξη. Κάποια από αυτά είναι το ποσοστό επικάλυψης λέξεων και το ποσοστό ομοιότητας της υποψήφιας φράσης με την

σελίδα προορισμού. Λαμβάνεται υπόψη επίσης, και η θέση της λέξης στην έκφραση της σελίδας προορισμού. Αυτά είναι κάποια δυαδικά χαρακτηριστικά, δηλαδή, πότε μια λέξη των υποψήφιων εκφράσεων παρουσιάζεται στον τίτλο της σελίδας προορισμού, ή στο κυρίως σώμα της σελίδας.

Βιβλιογραφία κεφαλαίου

[1] Panagiotis Papadimitriou. “Algorithms and Strategies for Web Advertising”. Stanford University, Dept. of Electrical Engineering. December 2011

[2] Sujith Ravi, Andrei Broder, Engeniy Gabrilovich, Vanja Josifovski, Sandeep Pandey, Bo Pang. “Automatic Generation of Bid Phrases for Online Advertising”. In *Proceedings of the third ACM international conference on Web search and data mining, Pages 341-350 (WSDM 2010).* New York City, USA. February 3-6, 2010.

Κεφάλαιο 3: Sponsored Search & Delivery

3.1 SSA (Sponsored Search Auctions with conflicts)

Οι διαφημιστές ανταγωνίζονται για m διαφημίσεις δίπλα στα αποτελέσματα αναζήτησης. Η συμμετοχή κάθε διαφημιστή στη δημοπρασία προσδιορίζεται από το πόσο σχετίζεται το αίτημα του χρήστη, με τις απαιτήσεις του διαφημιστή. Βέβαια, αν και αυτό το μοντέλο είναι απλό, του λείπουν κάποια χαρακτηριστικά έκφρασης. Για παράδειγμα, δεν υπάρχει περίπτωση από έναν διαφημιστή να εκφράσει την αξία ενός κλικ, και κατά συνέπεια η προσφορά του να εξαρτάται από την ταυτότητα των άλλων διαφημιστών που εμφανίζονται στα m slots (περιοχές-διαφημίσεις). Προς βελτίωση των παραπάνω, επιτρέπουμε στον διαφημιστή να προσδιορίσει ένα σύνολο από διαφημιστές όπου κανονικά δεν θα εμφανίζονταν στα μεγαλύτερα m slots (conflict advertisers). Ο διαφημιστής μπορεί να εκφράσει μια διαφορετική εκτίμηση για ένα κλικ, αν οι περιορισμοί του ικανοποιούνται, και μια διαφορετική προσφορά αν οι περιορισμοί του παραβιάζονται. Αυτό που κάνει αυτή τη διαδικασία ελκυστική είναι το γεγονός ότι μπορεί να επωφεληθούν όλες οι πλευρές. Οι διαφημιστές που δηλώνουν τις αντικρουόμενες διαφημίσεις μπορούν πιο εύκολα να χειριστούν το περιεχόμενο όταν οι διαφημίσεις τους κλικάρονται και έτσι μπορούν να αποφύγουν και το κόστος των ανεπιθύμητων διαφημίσεων.

3.1.1 Ορισμός προβλήματος

A = {α1,…,αj,…,an} : σύνολο n διαφημιστών

Κάθε διαφημιστής συμμετέχει στη δημοπρασία με μία μόνο διαφήμιση και θέλει η διαφήμισή του να κλικαριστεί από τους χρήστες. Ένας διαφημιστής aj συμμετέχει με την εκτίμηση του vj για ένα κλικ, με την προσφορά του bj, δηλαδή το μεγαλύτερο ποσό που θέλει να πληρώσει για ένα κλικ, και ένα σύνολο διαφημιστών Cj ⊆ A. Ένας διαφημιστής μπορεί να δηλώσει την αντίθεσή του ενάντια σε άλλους διαφημιστές, οι οποίοι χρησιμοποιούν συγκεκριμένες λέξεις κλειδιά στις διαφημίσεις τους. Οι διαφημιστές ενδιαφέρονται για τα κλικ των χρηστών, όμως η μηχανή αναζήτησης δεν μπορεί να δημοπρατήσει αυτά τα κλικ απευθείας. Η μηχανή αναζήτησης δημοπρατεί m ad slots S = {s1,…,si,…,sm}, όπου οι διαφημίσεις μπορούν να εμφανισθούν. Κάθε slot si έχει την δικιά του πιθανότητα κλικαρίσματος pi. Οι πιθανότητες υπολογίζονται μέσω του click through rate (CTR) και υποθέτουμε ότι μειώνονται καθώς ο δείκτης i αυξάνεται, δηλαδή pi ≥ pi+1, για κάθε si ϵS. Η μείωση της πιθανότητας κλικαρίσματος είναι αποτέλεσμα της θέσης των slots στη σελίδα αποτελεσμάτων, όπου τα slots εξαπλώνονται σε μια στήλη με ένα slot σε κάθε γραμμή. Όσο πιο κοντά είναι ένα slot, τόσο περισσότερο ελκύει την προσοχή ενός χρήστη και τόσο πιο πιθανό είναι να κλικαριστεί.

Δεδομένης της εκτίμησης του κλικ, δηλαδή το vj, και τις πιθανότητες κλικαρίσματος pi, η αναμενόμενη χρησιμότητα ενός διαφημιστή aj, ο οποίος κερδίζει το slot είναι vjpi. Αν ο διαφημιστής j δεν κερδίσει κανένα slot, τότε η χρησιμότητά του είναι 0. Η μηχανή αναζήτησης, πρέπει να δεσμεύσει τα δημοπρατούμενα slots για τους διαφημιστές και να τα κοστολογήσει αντίστοιχα. Έτσι, η μηχανή αναζήτησης χρειάζεται να ορίσει έναν κανόνα

δέσμευσης (allocation rule) και έναν κανόνα πληρωμής (payment rule). Ο σκοπός του κανόνα δέσμευσης είναι να μεγιστοποιήσει την ιδιότητα social welfare, δηλαδή να μεγιστοποιήσει το σύνολο των χρησιμοτήτων των διαφημιστών. Μια δυαδική μεταβλητή xij ϵ {0,1} δείχνει το πότε ένας διαφημιστής aj κερδίζει ένα slot si. Έτσι, το social welfare ορίζεται ως: $\sum_{j=1}^{n}\left(\sum_{\mathrm{i=1}}^{\mathrm{m}} \mathrm{v_j p_i x_{ij}}\right)$.

Επίσης, ο κανόνας δέσμευσης θα πρέπει να ικανοποιεί μια σειρά περιορισμών όπως φαίνεται στη συνέχεια:

$\sum_{j=1}^{n}\left(\sum_{\mathrm{i=1}}^{\mathrm{m}} \mathrm{v_j p_i x_{ij}}\right)$, $\left(\sum_{\mathrm{j=1}}^{n} \mathrm{x_{ij}}\right)$ ≤ 1 για κάθε si ϵ S, που δείχνει ότι κάθε slot δεσμεύεται για το πολύ έναν διαφημιστή, $\left(\sum_{\mathrm{i=1}}^{m} \mathrm{x_{ij}}\right)$ ≤ 1 για κάθε aj ϵ A ∧ ak ϵ Cj , που δείχνει ότι κάθε διαφημιστής κερδίζει το πολύ ένα slot και τέλος xij ϵ {0,1}, που εγγυάται ότι οι περιορισμοί των αντιθέσεων των διαφημιστών ικανοποιούνται. Όλα αυτά απαρτίζουν το λεγόμενο πρόβλημα δέσμευσης (allocation rule). Μια εφικτή δέσμευση μπορεί να είναι η δέσμευση slots στους διαφημιστές που ικανοποιούν όλους τους περιορισμούς ενός γραμμικού προβλήματος.

Για παράδειγμα, έστω ότι υπάρχουν τρείς διαφημιστές A = {a1,a2,a3} με τις εξής αντιθέσεις C1=C3=Ø και C2= {a1} και δύο ad slots. Εάν δώσουμε την δέσμευση των διαφημιστών σε slots σε μια ταξινομημένη λίστα, η δέσμευση [α1], δηλαδή ο διαφημιστής α1 κερδίζει το πρώτο slot, ή δέσμευση [α2,α3], δηλαδή ο διαφημιστής α2 κερδίζει το πρώτο slot και ο διαφημιστής α3 κερδίζει το δεύτερο slot, είναι κάποιες εφικτές δεσμεύσεις.

Όμως, στην υλοποίηση μας, κάθε διαφημιστής προσφέρει μόνο αν ικανοποιείται ο περιορισμός του που αφορά τις αντιθέσεις. Παρόλα αυτά, ένας διαφημιστής μπορεί να θέλει να προσδιορίσει δύο διαφορετικές προσφορές, μία για την περίπτωση που οι περιορισμοί αντιθέσεων ικανοποιούνται, και μία για την περίπτωση που παραβιάζονται. Στην υλοποίηση ακολουθούμε την εξής πορεία: για έναν διαφημιστή aj που θέλει να υποβάλει δύο προσφορές, κατασκευάζουμε μια εικονική διαφήμιση aj’ και σιγουρευόμαστε ότι το aj ϵ Cj’ και aj’ ϵ Cj. Έπειτα, ο διαφημιστής μπορεί να υποβάλλει δύο διαφορετικές προσφορές με διαφορετικά σύνολα αντιθέσεων. Δίνονται οι δεσμεύσεις x = {xij} και ένα διάνυσμα προσφοράς b = {bj}. Ο κανόνας πληρωμής ορίζει την τιμή pj(b,x) που ο διαφημιστής j χρειάζεται να πληρώσει. Μια επιθυμητή ιδιότητα για τον κανόνα πληρωμής είναι να επιτύχουμε αληθοφανείς προσφορές, δηλαδή να παρακινήσουμε τους διαφημιστές να αναφέρουν το πραγματικό τους κλικ καθώς και την αξία του (bj = vj). Η αποπληρωμή ενός διαφημιστή ορίζεται από την χρησιμότητα μείον την πληρωμή, δηλαδή $\left(\sum_{\mathrm{i=1}}^{\mathrm{m}} \mathrm{v_j p_i x_{ij}}\right)$ – pj(b,x). Έτσι, αν b, b’είναι δύο διανύσματα προσφορών τα οποία διαφέρουν μόνο στην j-στη θέση, με bj = vj και bj’ ≠ vj και x*, x’* είναι οι δεσμεύσεις από τον κανόνα δέσμευσης, τότε ο κανόνας πληρωμής διαβεβαιώνει ότι αν και μόνο αν για όλα τα b1,…,bj-1,bj+1,…,bn : $\left(\sum_{\mathrm{i=1}}^{\mathrm{m}} \mathrm{v_j p_{ij} x *_{ij}}\right)$ - pj(b,x*) ≥ $\left(\sum_{\mathrm{i=1}}^{\mathrm{m}} \mathrm{v_j p_{ij} x' *_{ij}}\right)$ - pj(b’,x’*).

3.1.2 Πολυπλοκότητα του προβλήματος

Παρουσιάζουμε τις αντιθέσεις του διαφημιστή και τις εκτιμήσεις του με ένα μη κατευθυνόμενο γράφημα G=(V,E). Κάθε κορυφή aj ϵ V αντιστοιχεί στον διαφημιστή aj, και το βάρος της κορυφής ισούται με την εκτίμηση κλικαρίσματος vj. Ο γράφος έχει μια ακμή

(aj,ak) αν ak ϵ Cj ή αν αj ϵ Ck. Παρατηρούμε ότι αφού ο γράφος είναι μη κατευθυνόμενος, δεν έχει σημασία αν οι διαφημιστές ak, aj δεν εμφανίζονται στα ad slots μαζί. Τελικά, αποδεικνύεται ότι είναι *NP-hard* να επιλυθεί το πρόβλημα δέσμευσης.

3.1.3 Ο αλγόριθμος του Debbassac

Algorithm 2 Debbassac: Depth-first-search branch-and-bound algorithm for sponsored search auctions with conflict constraints.

Input: set of advertisers A with conflict set C_j and valuations $v_j, \forall a_j \in A$, slots S with click probabilities $\rho_i, \forall s_i \in S$.
Output: optimal allocation x^* and social welfare w^*.

Global variable $w^* \leftarrow 0$ ▷ Optimal welfare
Global variable $X^* \leftarrow \emptyset$ ▷ Sparse representation of the optimal allocation
ALLOCREM($\emptyset, 0, A$)
$x^*_{ij} \leftarrow 0, \forall a_j \in A, s_i \in S$
for all $(s_i, a_j) \in X^*$ **do**
    $x^*_{ij} \leftarrow 1$
return x^*, w^*

function ALLOCREM($X_{\text{in}}, w_{\text{in}}, A_{rem}$)
Input invariants: $\{a_j \mid (s_i, a_j) \in X_{\text{in}}\} \sqcap A_{rem} = \emptyset$
    $i \leftarrow |X_{\text{in}}| + 1$
    ▷ Check for the end of recursion.
    if $i = m + 1 \vee A_{rem} = \emptyset$ **then**
        if $w_{\text{in}} > w^*$ **then**
            $w^* \leftarrow w_{\text{in}}$
            $X^* \leftarrow X_{\text{in}}$
        return
    ▷ Calculate w_j, h_j and $A_{j,rem}$ for $a_j \in A_{rem}$.
    ▷ A_{rem} sorted by v_j desc.
    for all $a_j \in A_{rem}$ **do**
        $w_j \leftarrow w_{\text{in}} + \rho_i v_j$
        $C'_j \leftarrow \{a_k \mid a_k \in A_{rem} \wedge a_j \in C_k\}$
        $L_j \leftarrow \{a_k \mid a_k \in A_{rem} \wedge v_k > v_j\}$
        $A_{j,rem} \leftarrow A_{rem} - (\{a_j\} \cup C_j \cup C'_j \cup L_j)$
        $h_j \leftarrow$ UPPERBOUND($m - i, A_{j,rem}$)
        if $(C_j \cup C'_j) \sqcap A_{rem} = \emptyset$ **then**
            break
    ▷ Explore allocations recursively.
    Remove un-examined advertisers from A_{rem}.
    Sort A_{rem} by $w_j + h_j$ desc.
    for all $a_j \in A_{rem}$ **do**
        if $w_j + h_j \leq w^*$ **then** ▷ Pruning condition
            break
        else
            ALLOCREM($X_{\text{in}} \cup \{(s_i, a_j)\}, w_j, A_{j,rem}$)

function UPPERBOUND(m_{rem}, A_{rem})
    $h \leftarrow 0$
    while $m_{rem} > 0 \wedge A_{rem} \neq \emptyset$ **do**
        $i \leftarrow m - m_{rem} + 1$
        $a_j \leftarrow \arg\max_{a_k \in A_{rem}} v_k$
        $h \leftarrow h + \rho_i v_j$
        $m_{rem} \leftarrow m_{rem} - 1$
        $A_{rem} \leftarrow A_{rem} - \{a_j\}$
    return h

1. Απαρίθμησε όλες τις δυνατές δεσμεύσεις των διαφημιστών στα ad slots. Λαμβάνουμε υπόψη μόνο δεσμεύσεις που είτε γεμίζουν όλα τα slots, είτε αφήνουν κάποια χαμηλά slots κενά. Οι δεσμεύσεις που αφήνουν ένα υψηλό slot κενό μπορούν να παραβλεφθούν, αφού μπορούμε πάντα να μετακινήσουμε τους διαφημιστές πιο ψηλά για να γεμίσουν ένα slot, αλλά και για να βελτιώσουν την ιδιότητα του social welfare που περιγράψαμε παραπάνω. Ο αριθμός των δυνατών δεσμεύσεων είναι n + $\binom{n}{2}$ + … + $\binom{n}{m}$. Το n είναι ο αριθμός των δεσμεύσεων που

γεμίζουν μόνο το υψηλό slot, το $\binom{n}{2}$ είναι ο αριθμός των δεσμεύσεων που γεμίζουν τα δύο υψηλά slots κοκ.

2. Αφαίρεσε δεσμεύσεις που δεν είναι εφικτές, δηλαδή δεσμεύσεις που περιέχουν διαφημιστές με αντιθέσεις ανάμεσά τους.
3. Βρες ανάμεσα στις υπόλοιπες δεσμεύσεις αυτή που φέρνει τη μεγαλύτερη τιμή στην ιδιότητα social welfare.

Αυτός ο αλγόριθμος είναι απαιτητικός, καθώς βρίσκει την καλύτερη δυνατή δέσμευση. Όμως, είναι ακριβός, αφού εξετάζει όλες τις πιθανές δεσμεύσεις. Στον αλγόριθμο του Debbassac δείχνεται πώς μπορούμε να περιορίσουμε το ψάξιμο ανάμεσα στο χώρο εφικτών δεσμεύσεων και πώς μπορούμε να βρούμε δεσμεύσεις που κερδίζουν άλλες που έχουν ήδη εξερευνηθεί.

Στο σημείο αυτό γίνεται προσπάθεια επεξήγησης των βημάτων του αλγόριθμου. Είσοδο του αλγόριθμου αποτελούν οι διαφημιστές A (μαζί με τις εκτιμήσεις των κλικ και τα σύνολα αντιθέσεων) και τα ad slots S (μαζί με τις πιθανότητες κλικαρίσματος). Ο αλγόριθμος ερευνά τον χώρο των εφικτών δεσμεύσεων και χρησιμοποιεί μια προσέγγιση για να υπολογιστεί ο χώρος έρευνας. Ο αλγόριθμος επιστρέφει την καλύτερη δέσμευση χ* και την αντίστοιχη social welfare w*. Στον ψευδοκώδικα χρησιμοποιείται η μεταβλητή X* σαν αναπαράσταση της καλύτερης δέσμευσης. Το X* είναι ένα σύνολο ζευγών όπως το (si,aj) που δείχνει ότι ο διαφημιστής aj κερδίζει το slot si.

Η έρευνα που αφορά το χώρο δέσμευσης υλοποιείται στην διαδικασία ALLOCREM(Xin,win,Arem). Η είσοδος της διαδικασίας είναι η δέσμευση διαφημιστών στις πρώτες i-1 slots Xin, η αντίστοιχη social welfare μέχρι στιγμής win, και ένα σύνολο διαφημιστών A_{rem} που μπορούν να γεμίσουν τις υπόλοιπες m – (i-1) slots. Η διαδικασία προσπαθεί να γεμίσει την i-στη slot με έναν από τους A_{rem} διαφημιστές και ξανακαλεί επαναληπτικά τον εαυτό της προκειμένου να γεμίσει τις υπόλοιπες m-i slots. Η διαδικασία δεν έχει επιστρεφόμενη τιμή, αλλά ανανεώνει τις καθολικές μεταβλητές w* και X*, όταν βρει δέσμευση που να προκαλεί υψηλότερη social welfare από το τρέχον μέγιστο.

Στις γραμμές 2-9 η διαδικασία ελέγχει πότε η δέσμευση εισόδου είναι συμπληρωμένη και συγκρίνει το social welfare με το τρέχον μέγιστο. Η δέσμευση είναι ολοκληρωμένη όταν, είτε τα m slots είναι γεμάτα, είτε αν δεν υπάρχουν άλλοι διαφημιστές που να απομένουν (A_{rem}=Ø) για να γεμίσουν τα εναπομείναντα slots. Αν η εισερχόμενη τιμή του social welfare win είναι μεγαλύτερη από το τρέχον μέγιστο, οι καθολικές μεταβλητές w* και X* αλλάζουν τιμή στις γραμμές 5-6. Η διαδικασία σταματά στη γραμμή 8. Αν η δέσμευση εισόδου δεν έχει ολοκληρωθεί, τότε ο αλγόριθμος εκτιμά τις δεσμεύσεις οι οποίες αυξάνουν το X_{in} με έναν διαφημιστή aj ϵ A_{rem} στο slot si(γραμμές 10-19). Πιο συγκεκριμένα, για κάθε τέτοια δέσμευση ο αλγόριθμος υπολογίζει:

α) την καινούργια τιμή social welfare wj. Εδώ ο υπολογισμός του wj είναι άμεσος.

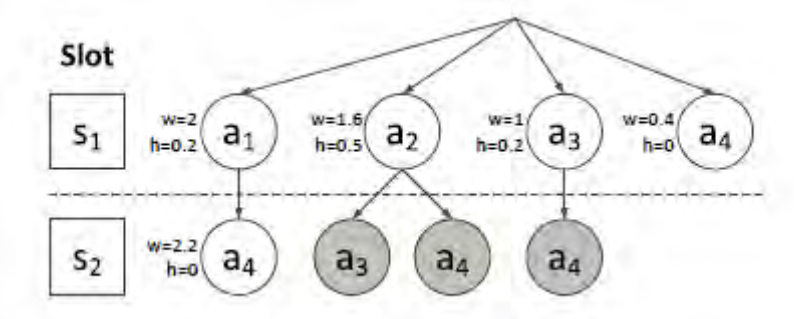
β) Το Aj,rem του διαφημιστή, το οποίο θα ληφθεί υπόψη για να γεμίσει τα υπόλοιπα m-i slots. Ο αλγόριθμος αφαιρεί από το σύνολο Arem, τον διαφημιστή aj, τις αντιθέσεις του Cj, τους διαφημιστές Cj’ που έχουν το aj στις αντιθέσεις τους, και τέλος τους διαφημιστές Lj, που έχουν μεγαλύτερη εκτίμηση από τον aj. Οι πρώτες τρείς αφαιρέσεις επιβεβαιώνουν ότι

ο αλγόριθμος θα ερευνήσει μόνο εφικτές δεσμεύσεις, καθώς προσπαθεί να γεμίσει τα υπόλοιπα m-i slots. Οι Lj διαφημιστές αφαιρούνται, γιατί δεν χρειάζεται να λαμβάνουμε υπόψη τον διαφημιστή ak με μεγαλύτερη εκτίμηση από τον aj για τα υπόλοιπα slots.

γ) Ένα άνω όριο hj του social welfare που τα υπόλοιπα slots μπορούν να συνεισφέρουν. Για να επιτύχουμε ένα άνω όριο υποθέτουμε ότι οι διαφημιστές με τις υψηλότερες m-i εκτιμήσεις κλικαρίσματος στο σύνολο Aj,rem θα κερδίσουν τα υπόλοιπα m-i slots. Αυτός ο υπολογισμός επιφέρει ένα άνω όριο, επειδή αγνοεί πιθανές αντιθέσεις ανάμεσα σε αυτούς τους m-i διαφημιστές.

Οι επαναλήψεις του αλγόριθμου σταματούν στις γραμμές 16-18, αν ο τρέχον διαφημιστής aj δεν έχει αντιθέσεις με τους υπόλοιπους διαφημιστές. Η έλλειψη αντιθέσεων επιβεβαιώνει ότι δεν επιβάλλεται κάποιος περιορισμός στη δέσμευση των υπόλοιπων slots. Επίσης, αφού οι διαφημιστές εξετάζονται με καθοδική σειρά των εκτιμόμενων κλικ, δέσμευση του slot si στον aj κυριαρχεί στην δέσμευση αυτού του slot σε οποιονδήποτε από τους ανεξέταστους διαφημιστές. Στις γραμμές 20-28, ο αλγόριθμος προσπαθεί να δεσμεύσει το slot si του διαφημιστή aj και έτσι να δεσμεύσει τα υπόλοιπα slots με επαναληπτικό τρόπο. Πρώτα πρώτα, ο αλγόριθμος ταξινομεί τους διαφημιστές του συνόλου Arem με το άθροισμα wj+hj προκειμένου να ερευνηθούν οι καλύτερα υποσχόμενες δεσμεύσεις πρώτα. Το άθροισμα wj+hj παρέχει ένα άνω όριο για το social welfare της οποιασδήποτε ολοκληρωμένης δέσμευσης, με τον διαφημιστή aj του i-στου slot και του πρώτου i-1 slot που έχει δεσμευθεί στο σύνολο Xin. Στις γραμμές 22-28, η διαδικασία καλεί τον εαυτό της για να ερευνήσει τη δέσμευση με το slot si του διαφημιστή aj, μόνο αν μια τέτοια δέσμευση μπορεί να προκαλέσει υψηλότερη τιμή social welfare από το τρέχον μέγιστο (γραμμή 23). Αν αυτή η συνθήκη δεν ικανοποιεί τη διαδικασία, σταματάει η επανάληψη και επιστρέφει. Σημειώστε ότι δε χρειάζεται να σκεφθούμε τους υπόλοιπους διαφημιστές, από τη στιγμή που έχουν ταξινομηθεί με αυξανόμενες τιμές των wj+hj. Αν η συνθήκη ισχύει, η διαδικασία καλεί επαναληπτικά τον εαυτό της με ανανεωμένη τιμή εισόδου δέσμευσης Xin U {(si,aj)}, η social welfare wj και υπολειπόμενων διαφημιστών Aj,rem.

Στο παρακάτω διάγραμμα φαίνεται πώς ο αλγόριθμος μπορεί να εφαρμοστεί στη δημοπρασία του παραδείγματος που τρέχει.



Σχήμα 5 : Παράδειγμα τρεξίματος του αλγορίθμου Debbassac [Πηγή: από Βιβλ. Κεφ., αναφορά 1]

Κάθε μονοπάτι του σχήματος από τη ρίζα στα φύλα, αντιπροσωπεύει μια ενδεχόμενη δέσμευση. Για παράδειγμα, το μονοπάτι α1->α4 αντιπροσωπεύει τη δέσμευση του slot s1 στο α1 και του slot s2 στο α4. Ο αλγόριθμος συμπεριλαμβάνει και τους τέσσερεις υποψήφιους διαφημιστές για να γεμίσουν το πρώτο slot. Αν ο διαφημιστής α1 επιλεγεί, τότε ο αλγόριθμος θα αναλογίζεται μόνο τον διαφημιστή α4 για το δεύτερο slot, αφού ο α4 έχει τον διαφημιστή α1 στα σύνολα αντιθέσεων του. Αν επιλεγεί ο διαφημιστής α2, τότε ο αλγόριθμος θα αναλογίζεται μόνο τους α3,α4 για το δεύτερο slot, αφού ο α2 έχει τον α1 στο σύνολο αντιθέσεων του. Αν επιλεγεί ο α3, τότε τον αλγόριθμο τον ενδιαφέρει μόνο ο α4 για το δεύτερο slot, αφού ο α4 έχει τον διαφημιστή α1 στο σύνολο αντιθέσεων του και ο α2 έχει υψηλότερη εκτίμηση κλικαρίσματος από τον α3. Τέλος, αν ο διαφημιστής α4 επιλεγεί, δε θα υπάρχει διαφημιστής να αναλογιστούμε για το δεύτερο slot, αφού οι υπόλοιποι διαφημιστές έχουν μεγαλύτερη εκτίμηση κλικαρίσματος από αυτόν. Ο αλγόριθμος επίσης υπολογίζει τιμές για το w και το h για κάθε κόμβο.

Στο παράδειγμα που θα δείξουμε υπάρχουν τέσσερεις διαφημιστές A= {α1,α2,α3,α4} με τιμές κλικαρίσματος v1=$10, v2= $8, v3= $5 και v4= $2 και έχουν τις ακόλουθες αντιθέσεις: C1=C4=Ø και C2=C3={a1}. Τα slots της δημοπρασίας είναι S={s1,s2} και οι πιθανότητες κλικαρίσματος είναι p1=0.2 και p2=0.1.

Έτσι, αν ο διαφημιστής α1 επιλεγεί για το πρώτο slot, τότε w=v1p1=$10 x 0.2 = $2 και h=v4p2= $2 x 0.1 = $0.2. Βασιζόμενοι στο άνω όριο της τιμής του social welfare w+h, ο αλγόριθμος θα επιλέξει να ερευνήσει πρώτα τη δέσμευση με το διαφημιστή α1 στο πρώτο slot. Έπειτα, ο αλγόριθμος δεσμεύει το δεύτερο slot στον διαφημιστή α4 και ανανεώνει το μέγιστο social welfare w*=$2.2. Αφού καμία από τις τιμές των άνω ορίων δεν είναι μεγαλύτερη από $2.2, ο αλγόριθμος θα τερματιστεί και ποτέ δεν θα ερευνήσει τους γκρι κόμβους του δέντρου αναζήτησης.

3.1.4 Συνδυαστική δημοπρασία (combinatorial auction)

Αυτός που κάνει την προσφορά σε μια συνδυαστική δημοπρασία μπορεί να τοποθετήσει την προσφορά του σε μια δέσμη, δηλαδή ένα συνδυασμό από δημοπρατούμενα αντικείμενα, και κερδίζει όλα ή τίποτα από τα αντικείμενα της δέσμης αυτής. Για παράδειγμα, αν τα δημοπρατούμενα αντικείμενα είναι {α,β,γ}, τότε μια προσφορά μπορεί να τοποθετηθεί στη δέσμη {α,β} και μια άλλη προσφορά στη δέσμη {β,γ}. Αφού οι δύο δέσμες έχουν το αντικείμενο β κοινό, μόνο ένας από αυτούς θα κερδίσει. Αν κερδίσει ο πρώτος, τότε το αντικείμενο γ δε θα δεικτοδοτηθεί, ενώ αν κερδίσει το δεύτερο αντικείμενο, τότε το αντικείμενο α δε θα δεικτοδοτηθεί. Χρησιμοποιούμε την έννοια των διαμοιραζόμενων αντικειμένων για να αναπαραστήσουμε τις αντιθέσεις μεταξύ των διαφημιστών. Επίσης, μαζί με την πραγματική διαφήμιση και το σύνολο των slots S, εισάγουμε και ένα σύνολο ψεύτικων slots D = {djk| aj ϵ A $\wedge$ (ak ϵ Cj $\vee$ aj ϵ Ck) $\wedge$ j $<$ k}. Το ψεύτικο slot αντιστοιχεί στην αντίθεση ανάμεσα στις αντιθέσεις των διαφημιστών aj και ak, δηλαδή το aj είναι το σύνολο αντιθέσεων ak. Σημειώνεται ότι η ανισότητα j < k στον ορισμό του συνόλου αποτρέπει τη δημιουργία δύο άχρηστων slots djk και dkj για την αντίθεση ανάμεσα στους διαφημιστές j και k. Αντί αυτού, ένα από τα δύο slots δημιουργείται.

Έτσι στο παράδειγμά μας, χρειάζεται να φτιάξουμε δύο ψεύτικα slots για να αναπαραστήσουμε τις αντιθέσεις του διαφημιστή: d12 για την αντίθεση του διαφημιστή α2

με τον α1, d13 για την αντίθεση του διαφημιστή α3 με τον α1. Το σύνολο των slots S και τα ψεύτικα slots D: R= S ∪ D. Με δεδομένα το αυξημένο σύνολο των slots, μπορούμε να μετατρέψουμε τις προσφορές των διαφημιστών, τα πραγματικά slots και τους περιορισμούς των αντιθέσεων του διαφημιστή aj, με τις ακόλουθες m δέσμες προσφορών: Bji = {si} ∪ Dj, i=1,…,m, όπου το σύνολο Dj των ψεύτικων slots που αντιστοιχεί σε αντιθέσεις που περιλαμβάνουν το αj είναι:

D = {djk| aj ϵ A ∧ djk ϵ D} ∪ {dkj |ak ϵ A ∧ dkj ϵ D}.

Οι δέσμες ότι οι τέσσερεις διαφημιστές προσφέρουν στο τρέχον παράδειγμα που αναλύουμε είναι:

$$B_{11} = \{s_1, d_{12}, d_{13}\}, \quad B_{12} = \{s_2, d_{12}, d_{13}\},$$
$$B_{21} = \{s_1, d_{12}\}, \quad B_{22} = \{s_2, d_{12}\},$$
$$B_{31} = \{s_1, d_{13}\}, \quad B_{32} = \{s_2, d_{13}\},$$
$$B_{41} = \{s_1\}, \quad B_{22} = \{s_2\}.$$

Κάθε δέσμη Bji περιέχει το πραγματικό slot si και τα ψεύτικα slots Dj. Έτσι, αν υπάρχει μια αντίθεση μεταξύ των διαφημιστών aj και ak και ο διαφημιστής aj κερδίσει μια δέσμη, λέει Bji, τότε ο διαφημιστής aj δεν κερδίζει μόνο το slot si, αλλά επίσης κερδίζει και το ψεύτικο slot djk. Το slot djk ανήκει και στον Dj και στον Dk. Αφού όλες οι δέσμες του διαφημιστή ak περιέχουν το ψεύτικο slot djk, ο διαφημιστής ak δεν μπορεί να κερδίσει καμία δέσμη, και συνεπώς, δεν μπορεί να κερδίσει κανένα πραγματικό slot.

Η εκτίμηση του διαφημιστή για τη δέσμη bji είναι ίση με την αναμενόμενη εκτίμηση από το κέρδος του πραγματικού slot si: vj(Bji) = vjpi. Το pi είναι η πιθανότητα κλικαρίσματος του slot si, και το vj είναι η εκτίμηση κλικαρίσματος του διαφημιστή aj. Ο διαφημιστής δεν εκτιμά τα ψεύτικα slots της δέσμης. Όμως, τα slots αυτά είναι σημαντικά για την ικανοποίηση των δικών του περιορισμών αντιθέσεων, αλλά και άλλων διαφημιστών. Η εκτίμηση δεσμών του διαφημιστή στο παράδειγμά μας είναι:

$$v_1(B_{11}) = \$20, \quad v_1(B_{12}) = \$10,$$
$$v_2(B_{21}) = \$16, \quad v_2(B_{22}) = \$8,$$
$$v_3(B_{31}) = \$10, \quad v_3(B_{32}) = \$5,$$
$$v_4(B_{41}) = \$2, \quad v_4(B_{42}) = \$1.$$

Όλοι οι διαφημιστές έχουν μηδενική εκτίμηση για κάθε άλλη δέσμη που δεν φαίνεται παραπάνω. Αν η δυαδική μεταβλητή xj(B) ϵ {0,1} δείχνει πότε ο διαφημιστής j κερδίζει την δέσμη B, ο κανόνας δέσμευσης που μεγιστοποιεί το social welfare πρέπει να λύσει το ακόλουθο γραμμικό πρόβλημα: $\sum_{j=1}^{n}\sum_{i=1}^{m} vj(Bji)xj(Bji)$, $\sum_{j=1}^{n}\sum_{B\epsilon B, r\epsilon B}^{m} xj(B) \leq 1$, ∀r ϵ R, $\sum_{i=1}^{m} xj(Bji) \leq 1$, ∀aj ϵ A, xj(B) ϵ {0,1} , όπου το σύνολο B = {Bji |aj ϵ A ∧ $si \epsilon S$} περιέχει όλες τις δέσμες slot που οι διαφημιστές κάνουν τις προσφορές τους. Αντικείμενο του προβλήματος είναι το social welfare να είναι ορισμένο στα πλαίσια των δεσμών δημοπρασίας. Το άθροισμα έχει το στοιχείο nm, όπου κάθε ένας από τους n διαφημιστές προσφέρει στις m δέσμες. Οι περιορισμοί εγγυώνται ότι σε κάθε slot r ϵ R, μπορεί να ανήκει

το πολύ σε μια δέσμη που είναι δεσμευμένη στους διαφημιστές. Με άλλα λόγια, οι δέσμες που είναι δεσμευμένες στους διαφημιστές δεν μπορούν να περιέχουν κοινά αντικείμενα. Ο αριθμός αυτών των ορίων είναι m +|D|, όπου το |D| είναι ίσο με τον αριθμό των αντιθέσεων του διαφημιστή. Τελικά, οι περιορισμοί των παραπάνω εξισώσεων εγγυώνται ότι κάθε διαφημιστής κερδίζει το λιγότερο μια δέσμη.

3.1.5 Ο κανόνας πληρωμής (Payment rule)

Ο πιο δημοφιλής κανόνας πληρωμής απαιτεί από τον διαφημιστή να κερδίσει το slot si για να πληρώσει την προσφορά του διαφημιστή ο οποίος κερδίζει s_{i+1}. Ο κανόνας αυτός, απαιτεί κάθε διαφημιστής να πληρώνει την μικρότερη προσφορά που χρειάζεται για να κερδίσει το slot. Δεν μπορούμε να εφαρμόσουμε αυτόν τον κανόνα στη διαδικασία SSA with conflicts, γιατί ο διαφημιστής πρέπει και να προσφέρει τις υψηλότερες προσφορές, αλλά επίσης πρέπει να προσφέρει και στις αντιθέσεις του. Προκειμένου να προσδιορίσουμε τις τιμές των διαφημιστών που κερδίζουν τα διαθέσιμα slots, χρησιμοποιούμε τον μηχανισμό Vickrey-Clarke-Groves. Μια γενική περιγραφή του μηχανισμού αυτού είναι η εξής: σε μια δημοπρασία οι διαφημιστές κάνουν γνωστή την αξία τους για τα αντικείμενα της δημοπρασίας (διαφημίσεις). Το σύστημα της δημοπρασίας αντιστοιχίζει τις διαφημίσεις με τρόπο ώστε να μεγιστοποιείται η ιδιότητα social welfare. Το σύστημα χρεώνει ατομικά κάθε διαφημιστή με το «κακό» που προκαλεί στους άλλους διαφημιστές – ανταγωνιστές του. Ο μηχανισμός αυτός μεγιστοποιεί το social welfare, όπου ο κάθε διαφημιστής πληρώνει το «κόστος ευκαιρίας», η παρουσία του οποίου εισάγεται σε όλους τους άλλους διαφημιστές.

Παρουσιάζουμε στη συνέχεια έναν τρόπο για το πώς εκτιμάται η τιμή σε κάθε διαφημιστή. Το uj*(A,S) δείχνει τη χρησιμότητα του διαφημιστή aj στην δέσμευση x* των slots S στους διαφημιστές A: uj*(A,S) = $\sum_{i=1}^{m} v_j p_i x^*_{ij}$. Στη συνέχεια, η τιμή για τον διαφημιστή aj, ισούται με τη διαφορά της συνολικής χρησιμότητας που οι άλλοι διαφημιστές λαμβάνουν, όταν ο aj συμμετέχει και όταν ο aj δεν συμμετέχει στη δημοπρασία: pj= $\frac{\sum_{k\neq j} u^*_k (A-\{j\},S)-\sum_{k\neq j} u^*_k(A,S)}{p_i}$.
Η διαφορά διαιρείται με το p_i, που είναι η πιθανότητα κλικαρίσματος του slot που έχει ανατεθεί στον aj, για να αποκτήσει την τιμή κάθε κλικ.

3.2 Βελτιστοποίηση για περιορισμό χρηματικών δαπανών των διαφημιστών (Optimizing budget constrained spend in search advertising)

Οι δημοπρασίες των μηχανών αναζήτησης έχουν τυπικά έναν σημαντικό περιορισμό διαφημιστών, με συγκεκριμένα χρήματα που μπορούν να διαθέσουν. Για παράδειγμα, αν επιτρέπεται να συμμετέχουν σε κάθε δημοπρασία, θα δαπανήσουν περισσότερα από αυτό που τους επιτρέπουν τα συνολικά χρήματά τους. Παρουσιάζουμε ένα σύστημα και αλγόριθμους προκειμένου να βελτιστοποιήσουμε αυτόν τον περιορισμό στις χρηματικές δαπάνες. Το πρόβλημα αυτό ονομάζεται optimized budget allocation.

Οι διαφημιστές χρησιμοποιούν φράσεις για να προσδιορίσουν το σύνολο των αιτημάτων χρηστών που ενδιαφέρονται, και προσφέρουν στο κόστος που διατίθενται να πληρώσουν για κάθε κλικ πάνω στη διαφήμιση. Επίσης, κάθε διαφημιστής προσδιορίζει ένα χρηματικό

ποσό, που είναι ένα άνω όριο χρημάτων που προτίθενται να πληρώσουν κάθε μέρα. Στο πρόβλημα λοιπόν στο οποίο αναφερόμαστε, ουσιαστικά δεσμεύονται κάποιοι διαφημιστές σε αιτήματα χρηστών, ώστε να ικανοποιούνται οι χρηματικοί περιορισμοί και ταυτόχρονα να βελτιστοποιείται κάποιος συγκεκριμένος στόχος. Θα μελετήσουμε δύο στόχους: την βελτίωση της ποιότητας των διαφημίσεων και την απόσβεση της χρηματικής επένδυσης (return on investment-ROI).

3.2.1 Ορισμός του προβλήματος

A: το σύνολο των διαφημιστών

Q: το σύνολο των ερωτήσεων

Κάθε διαφημιστής $\alpha \in A$ έχει ένα συγκεκριμένο χρηματικό ποσό Βα. Έστω το γράφημα G(A,Q,E), τέτοιο ώστε για τον διαφημιστή $\alpha \in A$ και $q \in Q$, η ακμή $(\alpha,q) \in E$ σημαίνει ότι η διαφήμιση του διαφημιστή α είναι συμβατή με το αίτημα q.

Ctr (a, q): η πιθανότητα ενός κλικ της διαφήμισης του α διαφημιστή στο αίτημα q. Αυτή η πιθανότητα δεν εξαρτάται από τη θέση της διαφήμισης.

Bid (a, q): τα χρήματα που ο διαφημιστής α διατίθεται να πληρώσει για κάθε κλικ

Όταν φτάνει ένα αίτημα q, οι συμβατές διαφημίσεις α για το q κατατάσσονται από τις τιμές των πιθανοτήτων bid, ctr. Συμβολίζοντας την j-στη διαφήμιση με τη σειρά σαν aj, το κόστος κάθε κλικαρίσματος aj ορίζεται ως: cpc(aj) = bid(aj+1,q)ctr(aj+1,q) / ctr(aj,q) και είναι γνωστό ως η γενικευμένη δεύτερη τιμή (generalized second price-GSP).

Τα : το έξοδο του διαφημιστή α, αν ο α συμμετέχει σε όλες τις δημοπρασίες, και η διαφήμιση α είναι συμβατή με τις λέξεις κλειδιά των διαφημίσεων. Αν το Τα > Βα, ο διαφημιστής είναι περιορισμένος χρηματικά, και η μηχανή αναζήτησης πρέπει να περιορίσει το σύνολο των δημοπρασιών στις οποίες συμμετέχει ο διαφημιστής.

Οι στόχοι που θέλουμε να επιτύχουμε, και η μηχανή αναζήτησης προσπαθεί να βελτιστοποιήσει είναι είτε η ποιότητα Τα των διαφημίσεων, δηλαδή να μεγιστοποιηθεί η πιθανότητα ctr, είτε να ελαττωθεί το κόστος κλικαρίσματος, είτε τέλος να αυξηθεί το λεγόμενο ROI (χρηματική επένδυση-return on investment).

Τελικά, το πρόβλημα του optimized budget allocation ορίζεται ως εξής: έχουμε πληροφορία για το παρελθόν μέσω ενός γράφου G'(A,Q',E') και γνωρίζοντας επίσης τις προσφορές και το CTR (click through rate) για κάθε ζευγάρι διαφημιστή και αιτήματος χρήστη, το χρηματικό ποσό Βα για κάθε διαφημιστή α, και μια συνάρτηση βελτιστοποίησης κάνουμε τα εξής: για κάθε αίτημα που φτάνει σε μια online ακολουθία αιτημάτων Q, αποφασίζεται ποιοι διαφημιστές θα συμμετέχουν στην δημοπρασία.

3.2.2 Αλγόριθμοι

Το πρώτο βήμα είναι να μην υπερξοδεύουμε, αλλά ούτε και να ξοδεύουμε λιγότερο. Ένας τρόπος για να γίνει αυτό είναι να αφήνουμε τους διαφημιστές να συμμετέχουν σε δημοπρασίες μέχρι να φτάσουν το χρηματικό τους ποσό, και μετά για την υπόλοιπη μέρα

να τους κάνουμε αδρανείς. Βέβαια, αυτό θα δημιουργήσει κίνηση στον διαφημιστή, και θα σταματήσει να υπάρχει ανταγωνισμός, όπως στο πρώτο κομμάτι της μέρας.

Μια δεύτερη απλή εκδοχή είναι το λεγόμενο VPT (Vanilla Probabilistic Throttling).

Για κάθε διαφημιστή α, ορίζουμε: Βα: το υπολειπόμενο ποσό της ημέρας.

Τα : το υπολειπόμενο μέγιστο ποσό που μπορεί να ξοδέψει για το υπόλοιπο της ημέρας, δηλαδή, το συνολικό έξοδο που θα έκανε αν ο διαφημιστής είχε απεριόριστο χρηματικό ποσό.

3.2.2.1 Βήματα αλγόριθμου VPT (Vanilla Probabilistic Throttling)

Για κάθε αίτημα q που φτάνει:

Για κάθε διαφημιστή α με περιορισμό στα χρήματα που δύναται να ξοδέψει:

Ρίξε νόμισμα με P[κορώνα] = Βα/Τα

Αν κορώνα, ο διαφημιστής α συμμετέχει στη δημοπρασία

Στη συνέχεια, παρουσιάζουμε κάποια βελτιστοποίηση για έναν ή περισσότερους από τους ακόλουθους στόχους: κατά μέσο όρο την ποιότητα των διαφημίσεων, εκπροσωπούμενη από το CTR, τα κλικ για κάθε δολάριο και άλλα, όπως το κέρδος του διαφημιστή, που χρησιμοποιεί τη διαφορά ανάμεσα στην προσφορά και στο κόστος κάθε κλικ σαν το εκτιμώμενο κέρδος. Για κάθε στόχο, με δεδομένη μια υποψήφια διαφήμιση i, υπολογίζουμε μια μετρική θ(i), η οποία εκτιμά τον στόχο. Ανάμεσα σε δύο στόχους και ανάλογα με την τιμή των μετρικών ποσοτήτων, θέλουμε να βρούμε τον στόχο με την υψηλότερη τιμή. Ο επόμενος αλγόριθμος λοιπόν, χρησιμοποιεί μια τρίτη είσοδο, το $R_{\theta,\alpha}$. Έχουμε επίσης τη συνάρτηση $F_{\theta,\alpha}(\mu)$, που είναι η εκτιμόμενη τιμή για το μέγιστο έξοδο Τα, για το οποίο το θ(i) ≤ μ. Με άλλα λόγια, το F είναι η συνάρτηση μάζας πιθανότητας του θ. Επίσης, έχουμε ότι $R_{\theta,\alpha} = 1- F_{\theta,\alpha}(\mu)$. Έτσι, ο βελτιστοποιημένος αλγόριθμος έχει ως εξής:

Για κάθε αίτημα χρήστη q που φτάνει:

Για κάθε διαφημιστή α:

Αν $R_{\theta,\alpha}(\theta(i)) \leq$ Βα / Τα, τότε ο διαφημιστής α συμμετέχει στη δημοπρασία

Παρατήρηση: ο αλγόριθμος ψάχνει για μια λύση, η οποία είναι δίκαια για κάθε διαφημιστή. Επίσης, παραθέτουμε τον παρακάτω πίνακα με τους πέντε στόχους που θέλουμε να επιτύχουμε:

	Objective	$\theta(i)$
OT-CTR	Ad quality	$\mathtt{ctr}_i$
OT-CLICKS	Clicks	$1/\mathtt{cpc}_i$
OT-PROFIT	Profit	$(\mathtt{bid}_i - \mathtt{cpc}_i)/\mathtt{cpc}_i$
OT-CONVERSIONS	Conversions	$\mathtt{cvr}_i\mathtt{val}_i/\mathtt{cpc}_i$
OT-CTR-PROFIT	Blend	$\frac{\mathtt{ctr}_i(\mathtt{bid}_i - \mathtt{cpc}_i)}{\mathtt{cpc}_i}$

Το ctr χρησιμοποιείται για να μετρηθεί η ποιότητα. Για τις επόμενες τρείς μετρικές προκειμένου να γίνουν κατανοητές θα πολλαπλασιάσουμε αριθμητή και παρονομαστή με το προβλεπόμενο CTR. Για παράδειγμα, στην μετρική OT-CLICKS, ο αριθμητής γίνεται ο αναμενόμενος αριθμός κλικαρισμάτων, και ο παρονομαστής γίνεται το αναμενόμενο κόστος, δηλαδή η συχνότητα είναι τα κλικς ανά δολάριο. Για την μετρική που αφορά το OT-PROFIT, υποθέτουμε ότι το bid_i είναι η αξία του διαφημιστή, το bid_i - cpc_i είναι το αναμενόμενο κέρδος του διαφημιστή, αν υπάρχει κάποιο κλικ σε αυτήν την εμφάνιση. Έτσι, η μετρική του OT-PROFIT είναι το αναμενόμενο κέρδος για κάθε δολάριο που ξοδεύεται. Για την μετρική που αφορά τα OT-CONVERSIONS έχουμε να κάνουμε με την αναμενόμενη τιμή μετατροπής ανά δολάριο, δοσμένης μια τιμής αντίθεσης σε μια εμφάνιση val_i, και ένα μοντέλο που προβλέπει τον ρυθμό αντίθεσης cvr_i. Τέλος, η τελευταία μετρική, απλά πολλαπλασιάζει την μετρική για το CTR και το κέρδος. Η διαίσθηση είναι ότι για διαφημιστές με μεγάλη ποικιλία στο CTR, αλλά όχι και τόσο στο κέρδος, ο αλγόριθμος θα εστιάσει στο CTR. Παρόμοια για διαφημιστές με περισσότερη διακύμανση στο κέρδος από το CTR, ο αλγόριθμος θα εστιάσει στο κέρδος. Έτσι, αν η μηχανή αναζήτησης νοιάζεται ταυτόχρονα για το CTR και για το κέρδος του διαφημιστή, η μετρική θα έχει καλύτερα αποτελέσματα από το να πάρουμε τον μέσο όρο των αποτελεσμάτων από ξεχωριστές μετρικές.

3.2.3 Σχεδιασμός του συστήματος

Στο σημείο αυτό θα περιγραφεί η υλοποίηση του αλγόριθμου, προκειμένου να γίνουν πιο κατανοητά τα βήματά του. Το σύστημα έχει τρία βασικά στοιχεία: τον υπολογισμό του Βα/Τα, τον υπολογισμό και την συμπίεση $R_{\theta,\alpha}$, και τέλος τη χρήση των υπολογισμών την ώρα της εξυπηρέτησης. Θα χρησιμοποιήσουμε την μετρική OT-CTR για να δείξουμε τις τεχνικές που χρησιμοποιήθηκαν και θα συζητήσουμε τις διαφορές ανάμεσα στο OT-CTR και σε άλλα στιγμιότυπα που συναντάμε.

Όσον αφορά το σχεδιασμό Βα/Τα, έχουμε να κάνουμε με τον υπολογισμό της πιθανότητας εμφάνισης ip, όπου το ip ορίζεται με το παραπάνω κλάσμα (Βα/Τα). Με άλλα λόγια, το ip είναι η πιθανότητα με την οποία θα πρέπει να επιτρέψουμε μια εμφάνιση του διαφημιστή σε μια δημοπρασία. Το ip υπολογίζεται με χρήση πληροφορίας της κίνησης του διαφημιστή από το παρελθόν και με χρήση του διαθέσιμου χρηματικού ποσού. Για τον υπολογισμό του $R_{\theta,\alpha}$ χρησιμοποιούμε δεδομένα του παρελθόντος για να υπολογίσουμε την αθροιστική συνάρτηση πιθανότητας $F_{\theta,\alpha}$. Το $R_{\theta,\alpha}$ είναι ένας τετριμμένος μετασχηματισμός του $F_{\theta,\alpha}$. Αυτό που θέλουμε είναι να συμπιέσουμε την πληροφορία στο R την ώρα της εξυπηρέτησης. Συμπιέζουμε το R σε ένα ιστόγραμμα H.

Τέλος, για το τρίτο στοιχείο που είναι η χρήση των υπολογισμών την ώρα της εξυπηρέτησης γίνεται η παρακάτω διαδικασία. Όταν φτάνει ένα αίτημα, χρειάζεται να προσδιορίσουμε για κάθε διαφημιστή α με περιορισμό στα χρήματα που θα δαπανήσει (budget constrained), πότε ο α συμμετέχει στη δημοπρασία. Η είσοδος αποτελείται από το ιστόγραμμα H, την τρέχουσα τιμή του ip και την τιμή του θ(i) για την τρέχουσα εμφάνιση i. Ας υποθέσουμε ότι το θ(i) βρίσκεται στο m. Το $H_b(m)$ δείχνει το άνω όριο του m και $H_r(m) = R(H_b(m))$. Έτσι, ο διαφημιστής συμμετέχει στη δημοπρασία με την ακόλουθη πιθανότητα:

$\begin{cases} 1, \ \alpha\nu \ H_r(m) \leq ip \\ 0, \ \alpha\nu \ H_r(m-1) \geq ip \\ \frac{ip - H_r(m-1)}{H_r(m) - H_r(m-1)}, \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases}$. Οι πρώτες δύο περιπτώσεις είναι ξεκάθαρες και προκύπτουν απευθείας από το γεγονός ότι στόχος είναι η εμφάνιση (ή μη) του διαφημιστή αν είναι (ή όχι) στο μέγιστο όριο δαπάνης ip. Στην τρίτη περίπτωση η εμφάνιση του διαφημιστή καθορίζεται από τον παραπάνω τύπο.

Στο σημείο αυτό, θα αναφέρουμε κάποια στοιχεία που αφορούν τη δράση μεταξύ διαφημιστών με περιορισμό στις χρηματικές τους δαπάνες. Έπειτα από ανάλυση αρχείων υπάρχει το εξής συμπέρασμα: οι διαφημίσεις με περιορισμό στις χρηματικές δαπάνες είναι περισσότερο πιθανόν να εμφανισθούν δίπλα σε διαφημίσεις που δεν έχουν τον παραπάνω περιορισμό. Στην περίπτωση αυτή, γίνεται χρήση επαναληπτικής τεχνικής για την βελτίωση της λειτουργίας των online αλγορίθμων. Η διαίσθηση πάντως όσον αφορά την επαναληπτική τεχνική είναι να δοκιμάσουμε επαναλήψεις της δημοπρασίας. Σε κάθε επανάληψη, υπολογίζουμε την μετρική θ για κάθε διαφημιστή με περιορισμό στις χρηματικές δαπάνες. Βασιζόμενοι στην τιμή της μετρικής αποφασίζεται για τον αν ο διαφημιστής θα συμμετέχει στην επόμενη επανάληψη.

3.3 Δεικτοδότηση και ανάκτηση στη διαδικασία του sponsored search

Η κύρια διαδικασία είναι να βρεθούν οι διαφημίσεις που είναι σχετικές με το αίτημα του χρήστη. Οι διαφημίσεις μπορούν να ανακαλυφθούν είτε με ακριβές ταίριασμα, όταν το αντικείμενο προσφοράς είναι ενδεικτικό στο αίτημα, είτε με πιο ακριβές ταίριασμα, όταν οι διαφημίσεις κατατάσσονται σαν έγγραφα. Αυτό είναι παρόμοια διαδικασία με την ανακάλυψη πληροφορίας (information retrieval-IR). Θα παρουσιαστούν λοιπόν τεχνικές για την κατάταξη των διαφημίσεων με χρήση και της IR τεχνικής. Επίσης, θα παρουσιαστεί και μια μέθοδος για ανακάλυψη διαφημίσεων που είναι περισσότερο σχετικές με το αίτημα του χρήστη, εκμεταλλευόμενοι την δομημένη φύση του σώματος της διαφήμισης. Αρχικά, μετατρέπουμε το σώμα της διαφήμισης σε ένα σύνολο από ιεραρχικά δομημένα έγγραφα, και μετά με χρήση της IR τεχνικής κατασκευάζεται ένας συμπαγές κατάλογος διαφημίσεων. Σε δεύτερο στάδιο, προτείνουμε διάφορες τεχνικές κατάταξης που πραγματεύονται την ιεραρχική δομή του σώματος μιας διαφήμισης προκειμένου να επιτύχουμε πιο αποδοτική κατάταξη διαφημίσεων.

3.3.1 Εύρεση και κατάταξη διαφημίσεων (Ad indexing and retrieval)

Το ζευγάρι <creative,term> είναι η μονάδα την οποία επεξεργαζόμαστε, αφού μαζί τα δύο στοιχεία αυτά αποτελούν μια διαφήμιση. Σκοπός είναι να παραχθεί μια ταξινομημένη λίστα σχετικών τέτοιων ζευγαριών. Τα δύο πεδία αυτά, το πρώτο (creative) και το δεύτερο το πεδίο των όρων (term) έχουν κάποιες υποκατηγορίες. Κάθε πεδίο creative αποτελείται από τα υποπεδία: τίτλος, περιγραφή και το URL. Κάθε ομάδα διαφημίσεων g αποτελείται από διάφορα πεδία <creative,term> που ομαδοποιούνται από τον διαφημιστή. Το σύνολο όλων των ομάδων διαφημίσεων συμβολίζεται με G και τα σύνολα όλων των <creative,term> συμβολίζονται με C και T αντίστοιχα. Το σύνολο των <creative,term> που σχετίζονται με μια συγκεκριμένη ομάδα διαφημίσεων συμβολίζεται με C^g και T^g αντίστοιχα. Άρα, ο συνολικός

αριθμός μοναδικών όρων που σχετίζονται με μια διαφήμιση είναι : |C| + |T| = |G|(| $\overline{C^g}$ |+| $\overline{T^g}$ |) . Το | $\overline{C^g}$ | και το | $\overline{T^g}$ | δείχνουν το μέσο όρο των creative και των όρων (term) σε κάθε ομάδα διαφημίσεων αντίστοιχα.

3.3.2 Συνάρτηση βαθμολόγησης και κατάταξης (Scoring Function and Ranking)

Όταν δίνεται ένα αίτημα q και μια μονάδα κατάταξης u, βαθμολογούμε κάθε μονάδα χρησιμοποιώντας το εξής μοντέλο: p(q/u)= $\prod_{w_i \in q} p(w_i/u)$, όπου το $p(w_i/u)$ υπολογίζεται χρησιμοποιώντας εξομάλυνση Dirichlet και άρα η συνάρτηση βαθμολόγησης γίνεται: $sc_q(u) \triangleq \sum_{w_i \in q} log \frac{tf_{w_i,u} + \mu \frac{tf_{w_i,C}}{\sum_{w_j \in C} tf_{w_j,C}}}{|u| + \mu}$. Το $tf_{w,k}$, είναι ο αριθμός των συμβάντων του όρου w, σε συγκεκριμένη μονάδα (k=u) ή σε ολόκληρη τη συλλογή (k=C). Το μ είναι μια ελεύθερη παράμετρος, η οποία ελέγχει την ποσότητα της εξομάλυνσης. Οι μονάδες κατάταξης μπαίνουν σε φθίνουσα σειρά, με βάση τη βαθμολογία τους. Ο χρήστης τυπικά βλέπει έναν περιορισμένο αριθμό αποτελεσμάτων σε απάντηση του αιτήματός του, και υποθέτουμε ότι μια λίστα των K μεγαλύτερων μονάδων [u1,…,uk] ανακτάται. Κάθε μονάδα u της λίστας αντιστοιχίζεται με έναν ανιχνευτή διαφημίσεων gId_u. Προκειμένου να υπάρχει ποικιλομορφία, μπορούμε να πούμε ότι το [$gIdu_1$,$gIdu_2$,…,$gIdu_k$] είναι μοναδικά, μειώνοντας έτσι τον αριθμό των διαφημίσεων που ανακτώνται από μια ομάδα διαφημίσεων σε μία.

3.3.3 Δομημένη κατάταξη διαφημίσεων (Structured Ad Indexes)

Οι στρατηγικές που θα παρουσιαστούν χρησιμοποιούν ομάδες διαφημίσεων και όρους προσφοράς.

3.3.3.1 Term Coupling Index (CTI)

Δεικτοδοτούμε μονάδες που περιλαμβάνουν ζευγάρια <creative,term> , <c,t>. Αυτή η προσέγγιση δεικτοδοτεί το καρτεσιανό παράγωγο των creative και των όρων (term) σε κάθε ομάδα διαφημίσεων. Είναι μια διαδικασία μονοεπίπεδη, αφού οι μονάδες δεικτοδότησης αντιστοιχούν σε διαφημίσεις. Η κατάταξη αυτή, δηλαδή η CTI, χρειάζεται να φιλτράρεται μόνο με το gId_t, αφού ανακαλύπτουν μια μοναδική διαφήμιση για κάθε ομάδα διαφημίσεων. Παρακάτω παρουσιάζεται ο ψευδοκώδικας για τον αλγόριθμο:

Algorithm 1 CTRank (**CTInd** index)

1: $\langle \mathbf{c}, \mathbf{t} \rangle \leftarrow [u_{(1)}, \ldots, u_{(K)}] \subseteq \mathbf{C} \times \mathbf{T}$
2: FILTER $(\mathbf{c}, \mathbf{t})$ BY gId_t
3: **return** $\langle \mathbf{c}, \mathbf{t} \rangle$

Ο μέσος όρος των μονάδων δεικτοδότησης για κάθε ομάδα διαφημίσεων είναι αντικείμενο παραγωγής των |$\overline{C^g}$ | και | $\overline{T^g}$|, που δείχνουν το μέσο όρο των creative και των όρων σε κάθε ομάδα διαφημίσεων αντίστοιχα. Κάθε τέτοια μονάδα περιέχει δύο πεδία το creative και το term. Έτσι, ο αναμενόμενος αριθμός πεδίων που δεικτοδοτούνται είναι 2|G||$\overline{C^g}$ || $\overline{T^g}$|.

3.3.3.2 Creative Coupling Index (CrtvInd)

Η μονάδα δεικτοδότησης σε αυτήν την κατάταξη είναι ένα μοναδικό c συνοδευόμενο με όλους τους όρους προσφοράς που σχετίζονται με κάθε ομάδα διαφήμισης gIdc. Είναι μια μικρότερη κατάταξη από το CTI, αφού πλέον δεν υπολογίζει κάποιο καρτεσιανό παράγωγο κάθε πεδίου. Αρχικά, με αυτή τη μέθοδο ανακτάμε μια ταξινομημένη λίστα του πεδίου creative, η οποία φιλτράρεται μέσω του gIdc και μετά βρίσκουμε τον όρο με την μεγαλύτερη βαθμολογία, η οποία σχετίζεται με κάθε όρο creative στην κατάταξη. Ο παρακάτω αλγόριθμος δείχνει την διαδικασία:

Algorithm 2 CrtvRank (CrtvInd index)

$\mathbf{c} \leftarrow [u_{(1)}, \ldots, u_{(K)}] \subseteq \mathbf{C}$
FILTER $\mathbf{c}$ BY gId_c
$\mathbf{t} \leftarrow [\text{argmax}_{t \in gId_c}\, sc_q(t) : c \in \mathbf{c}]$
return $\langle \mathbf{c}, \mathbf{t} \rangle$

Ο δείκτης CrtvInd, και κάθε πεδίο creative που κατατάσσουμε, κατά μέσο όρο, 1 + | $\overline{T^g}$| πεδία όρων, και συνολικά υπάρχουν |G||$\overline{C^g}$ | creatives στη συλλογή. Ο αναμενόμενος αριθμός των δεικτοδοτούμενων πεδίων είναι λοιπόν |G||$\overline{C^g}$ |(1+ | $\overline{T^g}$|).

3.3.3.3 Ad Group Coupling Index (AdGrpInd)

Στην προσέγγιση αυτή, η μονάδα δεικτοδότησης είναι η ίδια η ομάδα διαφημίσεων. Προκειμένου να βρούμε ένα ζευγάρι <creative,term>, πρώτα ανακαλύπτουμε μια ταξινομημένη λίστα από ομάδες διαφημίσεων. Μετά, για κάθε ομάδα, ένας όρος creative και ένας όρος με την υψηλότερη βαθμολογία ανακαλύπτονται. Από τη στιγμή που οι βαθμολογίες αυτές είναι ανεξάρτητες, και υποθέτοντας ότι η συνάρτηση βαθμολογίας είναι μονοτονική, το ζευγάρι <c,t> είναι το ζευγάρι με την υψηλότερη βαθμολογία στην ομάδα διαφημίσεων. Ο αλγόριθμος παρουσιάζεται παρακάτω:

Algorithm 3 AdGrpRank (AdGrpInd index)

$\mathbf{g} \leftarrow [u_{(1)}, \ldots, u_{(K)}] \subseteq \mathbf{G}$
$\mathbf{c} \leftarrow [\text{argmax}_{c \in gId_g}\, sc_q(c) : g \in \mathbf{g}]$
$\mathbf{t} \leftarrow [\text{argmax}_{t \in gId_g}\, sc_q(t) : g \in \mathbf{g}]$
return $\langle \mathbf{g}, \mathbf{c}, \mathbf{t} \rangle$

Σημειώνεται ότι ο δείκτης AdGrpInd είναι ο μόνος τύπος δείκτη που δεν έχει διπλά πεδία. Ο αριθμός των μονάδων δεικτοδότησης είναι ο ίδιος με τον αριθμό των ομάδων διαφημίσεων, |G|, και για κάθε ομάδα διαφήμισης υπάρχουν κατά μέσο όρο| $\overline{C^g}$ |+| $\overline{T^g}$ | πεδία. Έτσι, ο αριθμός των δεικτοδοτούμενων πεδίων είναι |G|(| $\overline{C^g}$ |+| $\overline{T^g}$ |). Αυτό το σχήμα είναι το πιο συμπαγές και μειώνει την ανάγκη να φιλτράρουμε τα αποτελέσματα με τον αναγνωριστή ομάδας διαφήμισης gIdu, αφού κάθε διαφήμιση κατασκευάζεται από ξεχωριστή ομάδα διαφήμισης.

3.3.3.4 Μέθοδος δομημένης επαναβαθμολόγησης (Structured reranking)

Θα δείξουμε ότι η βασική κατάταξη δεν είναι καμιά φορά επαρκής για να παράγουμε τα καλύτερα αποτελέσματα. Εδώ λαμβάνουμε υπόψη ότι δεδομένου του ζεύγους <creative,term> και της αντίστοιχης ομάδας διαφημίσεων, η μέθοδος αυτή εκτελεί έναν αρχικό γύρο εξερεύνησης, χρησιμοποιώντας το AdGrpRank. Έπειτα, η μέθοδος αξιολογεί ξανά τις διαφημίσεις που βρέθηκαν αρχικά με χρήση ενός γραμμικού μοντέλου: $rSc_q(g,c,t) \triangleq \sum_{i=1}^{n} \lambda_i f_i(g,c,t)$, όπου η συνάρτηση f είναι μια συνάρτηση χαρακτηριστικών, το λ_i είναι τα βάρη που δίνονται σε κάθε συνάρτηση και το n είναι ο αριθμός αυτών των συναρτήσεων που χρησιμοποιούνται για την επαναξιολόγηση. Παρακάτω παρουσιάζεται ο αλγόριθμος:

Algorithm 4 StructRank (**AdGrpInd** index)

1: $\langle \mathbf{g}, \mathbf{c}, \mathbf{t} \rangle \leftarrow$ **AdGrpRank**
2: SORT DESC $\langle \mathbf{g}, \mathbf{c}, \mathbf{t} \rangle$ BY $rSc_q(g,c,t)$
3: **return** $\langle \mathbf{g}, \mathbf{c}, \mathbf{t} \rangle$

Η επιλογή των χαρακτηριστικών που χρησιμοποιούνται παίζουν σημαντικό ρόλο στην τελική συμπεριφορά του αλγόριθμου. Με το να μειώσουμε την επιλογή των χαρακτηριστικών με χρήση βαθμολογίας μόνο στα πεδία οδήγησε σε ένα μικτό μοντέλο που συνοψίζεται στην ακόλουθη εξίσωση: $sc_q(g,c,t) = \lambda sc_q(uc,l) + (1-\lambda)sc_q(g)$. Το uc,l είναι μια μονάδα <creative,term>, <c,t> και το $sc_q()$ ορίζεται παραπάνω.

3.4 Χρήση σελίδων προορισμού για την επιλογή διαφημίσεων

Στην παράγραφο αυτή, θα περιγραφούν δύο είδη τεχνικών για την επιλογή χρήσιμων περιοχών των σελίδων προορισμού (landing pages). Είναι η μέθοδος out-of-context και in-context. Η μέθοδος out-of-context επιλέγει περιοχές της σελίδας αναλύοντας το περιεχόμενο, χωρίς να λαμβάνει υπόψη της τη διαφήμιση που σχετίζεται με τη σελίδα προορισμού. Στη μέθοδο in-context, γίνεται χρήση του περιεχομένου της διαφήμισης και βοηθά στο να αναγνωριστούν οι περιοχές της σελίδας προορισμού που θα πρέπει να χρησιμοποιηθούν από τη μηχανή αναζήτησης διαφημίσεων. Οι τεχνικές αυτές, μειώνουν το μέγεθος των σελίδων προορισμού, ενώ ταυτόχρονα διατηρούν ή βελτιώνουν την διαδικασία της αναζήτησης διαφημίσεων, ενάντια στη μέθοδο που χρησιμοποιεί ολόκληρη τη σελίδα προορισμού. Επίσης, θα παρουσιαστεί ένας αλγόριθμος που χρησιμοποιείται για να εμπλουτίσει το περιεχόμενο της διαφήμισης και έτσι να βελτιωθεί περαιτέρω η επιλογή διαφημίσεων.

3.4.1 Επιλογή όρων in – context

3.4.1.1 Ανακάλυψη σχετικών περιοχών – αλγόριθμος 1

Στο σημείο αυτό θα παρουσιαστεί η διαδικασία του να επιλεγούν σχετικές μεταξύ τους περιοχές στη σελίδα προορισμού (landing page) με περιεχόμενο που να σκοπεύει να παρουσιάσει η διαφήμιση. Τα βήματα που ακολουθούν δείχνουν τη διαδικασία επιλογής

όρων από τη σελίδα προορισμού βρίσκοντας τις σχετικές περιοχές με βάση το περιεχόμενο της διαφήμισης.

a := TF-logIDF representation of an ad
CR := {[-5,+5] landing page text around any word in a}
$RR := \emptyset$
For each candidate region $r_i \in CR$,
  If $cosine\_similarity(a, r_i) > \delta$,
  Then $RR \leftarrow RR \cup r_i$,
Return RR as relevant regions for the given ad

Χρησιμοποιείται το tf-idf μοντέλο για να αναπαραστήσει τις διαφημίσεις. Το κείμενο που χρησιμοποιείται στη διαφήμιση αποτελείται από τρία μέρη: τον τίτλο, μια μικρή περιγραφή και μια λέξη προσφοράς (bid phrase).

Για παράδειγμα, στην διαφήμιση που ακολουθεί όροι που επαναλαμβάνονται σε διαφορετικά στοιχεία, όπως η λέξη «machine» και «learning», φαίνεται πως είναι περισσότερο σημαντικές από λέξεις που επαναλαμβάνονται μέσα στο κείμενο, αλλά σε μόνο ένα στοιχείο.

TITLE	Machine Learning
BID PHRASE	machine learning
SHORT DESCRIPTION	Compare Prices in 101+ stores. Find cheap book prices every time.

Για τον υπολογισμό της συχνότητας εμφάνισης των όρων, μετράμε τον αριθμό των στοιχείων που κάποιος όρος εμφανίζεται. Με παρόμοιο τρόπο συμπεριφερόμαστε σε κάθε στοιχείο διαφήμισης σαν ένα ξεχωριστό έγγραφο. Το διάνυσμα που προκύπτει ονομάζεται ad vector. Στη συνέχεια, εντοπίζουμε υποψήφιες περιοχές της σελίδας προορισμού στο πλαίσιο του παραπάνω διανύσματος. Για κάθε λέξη της σελίδας προορισμού που επίσης εμφανίζεται και στο διάνυσμα, λαμβάνουμε υπόψη το παράθυρο από [-5,5] σαν μια υποψήφια περιοχή. Για κάθε τέτοια περιοχή υπολογίζουμε κατά πόσο μοιάζουν μεταξύ τους η υποψήφια περιοχή και το διάνυσμα. Ανακατεύουμε όλες τις υποψήφιες περιοχές των οποίων τα σκορ ομοιότητας είναι πάνω από ένα όριο δ. Οι περιοχές που καταλήγουμε είναι σχετικές περιοχές της δοσμένης διαφήμισης. Προκειμένου να αποφύγουμε το πρόβλημα κάποιες περιοχές να μην επιλεχθούν σαν σχετικές, εξαιτίας του μη ταιριάσματος της διαφήμισης και της σελίδας προορισμού, εισάγουμε έναν αλγόριθμο που επεκτείνει το διάνυσμα σε ένα με πλουσιότερο περιεχόμενο. Θα γίνει χρήση διανύσματος για όλες τις λέξεις μιας διαφήμισης και έτσι θα κατασκευαστεί ένα σημασιολογικό διάνυσμα όπου θα αναπαριστά τη σημασιολογική ερμηνεία του σκοπού της διαφήμισης. Ο αλγόριθμος αυτός θα παρουσιαστεί σε επόμενη παράγραφο.

3.4.1.2 Κατασκευή διανύσματος από το σώμα της διαφήμισης

Ορίζουμε τη μέτρηση για ένα ζεύγος λέξεων u και w ως τον αριθμό των ψευδοεγγράφων που εμφανίζονται: $cooc_{cnt}(u,w) = |\{d/u \in d \wedge w \in d\}|$. Παραλείποντας λέξεις που δεν εμφανίζονται συχνά καταλήγουμε στο εξής διάνυσμα για κάθε λέξη u: $cooc_{vec}(u) = \{w/cooc_{cnt}(u,w) > 0\}$. Για όλα τα $w \in cooc_{vec}(u)$ υπολογίσαμε το λεγόμενο PMI (point-wise mutual information). Το PMI ορίζεται ως : $PMI(u,w) = log \frac{\frac{c_{uw}}{N}}{\frac{\sum_{i=1}^{n} c_{iw}}{N} \times \frac{\sum_{j=1}^{n} c_{uj}}{N}}$, όπου το c_{uw} είναι ο αριθμός των φορών το u και w να εμφανίζονται μαζί, n είναι ο αριθμός των μοναδικών λέξεων και N είναι ο συνολικός αριθμός εμφανίσεων.

Για παράδειγμα, έχουμε το εξής διάνυσμα επανεμφάνισης (co-occurrence vector):

CENTRAL WORD	CO-OCCURRING WORDS(PMI)
mattress	futon(6.4), king(2.95), pillow(4.92) queen(5.64), shopping(2.2), brand(2.5) tempur-pedic(6.66), bunk(5.28), mite(5.79) serta(7.64), sealy(7.79), visco(7.75) platform(4.74), products(1.94), store(2.44) cover(4.1), outlet(3.46), directory(2.4) savings(1.37), topper(5.71), allergen(6.63)

Το u = «mattress» που δίνεται παραπάνω, δηλαδή είναι η κεντρική λέξη που μας ενδιαφέρει. Η βαθμολογία PMI που προκύπτει για κάθε λέξη φαίνεται στην παρένθεση. Η βαθμολογία αυτή δείχνει πόσο κατατοπιστική είναι μια λέξη επανεμφάνισης (co-occurring word) για τη λέξη u. Έτσι, λέξεις με μεγαλύτερο PMI όπως η λέξη «futon», «tempur-pedic», «serta», δίνουν σε γενικές γραμμές περισσότερη πληροφορία, από λέξεις με μικρότερη βαθμολογία PMI όπως π.χ. «shopping», «products», «savings».

3.4.1.3 Σύνθεση σημασιολογικού διανύσματος

Σκοπός μας είναι πώς θα συνδυάσουμε σε ένα μοναδικό διάνυσμα τα παραπάνω. Το {ui} είναι το σύνολο των λέξεων που αντιπροσωπεύουν μια διαφήμιση και V= {v1,v2,…,vn} είναι το σύνολο των PMI βασισμένο σε διανύσματα της διαφήμισης αυτής, όπως το $vi = \{vij | j \in cooc_{vec}(ui)\}$ και το vij είναι το σύνολο των τιμών του PMI ανάμεσα στη λέξη ui και j. Προκειμένου να συνθέσουμε αυτά τα διανύσματα σε ένα και μοναδικό διάνυσμα, το ονομαζόμενο compositional semantic vector (csv) για μια διαφήμιση έχουμε: csv = f(v1,…,vn). Επίσης, ερευνούμε διάφορα διανύσματα σύνθεσης προκειμένου να συνθέσουμε ένα σημασιολογικό διάνυσμα για μια διαφήμιση. Ένα από αυτά είναι το: $csv_j = \sum_i v_{ij}$, όπου το csv_j και το v_{ij} είναι το j-στο στοιχείο του διανύσματος csv και του vi αντίστοιχα.

Υπάρχουν κάποιες λέξεις που κουβαλάνε πληροφορία, αλλά έχουν μικρότερη συνεισφορά όταν συνθέτουν τη σημασιολογική ερμηνεία ολόκληρης της διαφήμισης. Για παράδειγμα, οι λέξεις «Compare» και «Find» στο παράδειγμα της παραγράφου 3.4.1.1 δεν δίνουν τόση πληροφορία, όπως π.χ. η λέξη «book» στο διαχωρισμό της συγκεκριμένης διαφήμισης από άλλες. Στην πράξη, οι λέξεις «Compare» ή «Find» μπορεί να εμφανίζονται σε σχεδόν όλη τη διαφήμιση, άσχετα αν το αντικείμενο της διαφήμισης είναι τελείως διαφορετικό. Παρόλα αυτά, τέτοιες λέξεις με αυτού του είδους την πληροφορία θα πρέπει να δίνουν σχετικά

μικρότερη συνεισφορά, όταν δημιουργείται η εννοιολογική ερμηνεία ολόκληρης της διαφήμισης. Έτσι, βάζουμε ένα βάρος σε κάθε διάνυσμα με βάση το μέσο όρο της βαθμολογίας PMI και έτσι οι παραπάνω εξίσωση γίνεται: $csv_j = \sum_i avgPMI(ui)v_{ij}$.

3.4.1.4 Ανακάλυψη σχετικών περιοχών – αλγόριθμος 2

Η παρακάτω διαδικασία περιγράφει το πώς θα βρεθούν οι σχετικές περιοχές με το εμπλουτισμένο περιεχόμενο.

a := TF-logIDF representation of an ad
csv := compositional semantic vector of an ad
c := TF-logIDF representation of N best entries of csv
CR_+ := {[-5,+5] landing page text around any word in a or c}
$RR_+ := \emptyset$
For each candidate region $r_i \in CR_+$,
  If $cosine\_similarity(a, r_i) + cosine\_similarity(c, r_i) > \delta_+$,
  Then $RR_+ \leftarrow RR_+ \cup r_i$,
Return RR_+ as relevant regions for the given ad

Αρχικά, αναπαριστούμε το περιεχόμενο της διαφήμισης με ένα διάνυσμα (ad vector). Στη συνέχεια, υπολογίζουμε τη σύνθεση του σημασιολογικού διανύσματος (compositional semantic vector-csv) της διαφήμισης. Κρατάμε τις μεγαλύτερες N εισόδους με την υψηλότερη βαθμολογία στο csv, προκειμένου να κρατήσουμε το μέγεθος του csv όμοιο με αυτό του διανύσματος διαφήμισης. Αυτό γίνεται για να επιβεβαιώσουμε ότι το εκτεταμένο διάνυσμα δε θα είναι μεγαλύτερο από το csv. Πριν συνδυαστούν το csv με το διάνυσμα διαφήμισης, υπολογίζουμε το σκορ tf-idf για τα διανύσματα διαφήμισης. Οι όροι που δεν εμφανίζονται στη διαφήμιση παίρνουν tf σκορ 1. Όταν υπολογιστεί το σημασιολογικό διάνυσμα c, ορίζουμε τις υποψήφιες περιοχές στη σελίδα προορισμού με τρόπο που περιγράψαμε παραπάνω. Για κάθε λέξη στη σελίδα προορισμού, που επίσης εμφανίζεται στο α ή το c, λαμβάνουμε υπόψη το παράθυρο κειμένου από [-5,5], σαν μια υποψήφια περιοχή. Για κάθε τέτοια περιοχή, υπολογίζεται η ομοιότητα ανάμεσα στην υποψήφια περιοχή και το διάνυσμα διαφήμισης, αλλά και η ομοιότητα ανάμεσα στην υποψήφια περιοχή και το σύνθετο σημασιολογικό διάνυσμα. Αν το άθροισμα της παραπάνω βαθμολογίας είναι μεγαλύτερο από ένα όριο δ_+ , τότε η υποψήφια περιοχή επιλέγεται σα σχετική περιοχή.

3.4.2 Επιλογή όρων out-of-context

Στο σημείο αυτό, ερευνούμε αλγορίθμους που βρίσκουν μια περιληπτική αναπαράσταση της σελίδας προορισμού, χωρίς να λαμβάνεται υπόψη η διαφήμιση που σχετίζεται με τη συγκεκριμένη σελίδα προορισμού.

3.4.2.1 Οι πρώτες N μοναδικές λέξεις

Μια δημοφιλής στρατηγική είναι να παίρνουμε σαν περίληψη την υψηλότερη παρτίδα μιας σελίδας προορισμού. Παίρνουμε έως N μοναδικές λέξεις που εμφανίζονται πρώτα στη σελίδα προορισμού.

3.4.2.2 Οι καλύτερες Ν μοναδικές λέξεις

Εδώ, λαμβάνουμε υπόψη Ν λέξεις που είναι οι πιο αντιπροσωπευτικές της σελίδας προορισμού. Χρησιμοποιούμε τα TF-IDF βάρη προκειμένου να εξάγουμε αυτές τις λέξεις.

3.4.2.3 Όλες οι λέξεις

Τέλος, δοκιμάζουμε και με όλες τις λέξεις της σελίδας προορισμού σαν μια ακραία περίπτωση. Αυτή η εκδοχή βέβαια δεν είναι και τόσο ελκυστική, γιατί δεν μειώνει τον όγκο πληροφορίας που χρειάζεται να δεικτοδοτηθεί.

Βιβλιογραφία κεφαλαίου

[1] Panagiotis Papadimitriou. “Algorithms and Strategies for Web Advertising”. Stanford University, Dept. of Electrical Engineering. December 2011

[2] Chinmay Karande, Aranyak Mehta, Ramakrishnan Srikant. “Optimizing Budget Constrained Spend in Search Advertising”. In *Proceedings of the sixth ACM international conference on Web search and data mining, Pages 697-706 (WSDM 2013).* Rome, Italy. February 4-8, 2013.

[3] Michael Bendersky, Evgeniy Gabrilovich, Vanja Josifovski, Donald Metzler. “The anatomy of an Ad: Structured Indexing and Retrieval for Sponsored Search”. In *Proceedings of the 19th international conference on the World Wide Web, (WWW 2010)*. Raleigh, North Carolina, USA. April 26-30, 2010.

[4] Yejin Choi, Marcus Fontoura, Evgeniy Gabrilovich, Vanja Josifovski, Mauricio Mediano, Bo Pang. “Using Landing Pages for Sponsored Search Ad Selection”. In *Proceedings of the 19th international conference on the World Wide Web, (WWW 2010)*. Raleigh, North Carolina, USA. April 26-30, 2010.

Κεφάλαιο 4: Sponsored Search & Response

4.1 Μοντέλο κλικαρίσματος και εφαρμογές (A novel click model and its applications to online advertising)

Εδώ θα παρουσιαστεί ένα μοντέλο κλικαρίσματος το οποίο ονομάζεται General Click Model (GCM) για να μάθουμε και να προβλέπουμε τη συμπεριφορά του χρήστη στα URLs που εμφανίζονται στη μηχανή αναζήτησης. Αρχικά, το μοντέλο αυτό βασίζεται σε πολλαπλά χαρακτηριστικά, και η επιρροή της τιμής κάθε διαφορετικού χαρακτηριστικού μπορεί να μετρηθεί με βάση τον πιθανοτικό νόμο του Bayes. Σε δεύτερο στάδιο, ενώ οι περισσότερες πιο παλιές τεχνικές λαμβάνουν υπόψη τη θέση και την ταυτότητα των URL, η τεχνική του GCM στρέφεται περισσότερο σε χαρακτηριστικά που αφορούν τον τομέα αυτό. Το μοντέλο επίσης, είναι ικανό να μοντελοποιεί τα τελευταία αιτήματα, επειδή βασίζεται σε τιμές χαρακτηριστικών οι οποίες μοιράζονται ανάμεσα στα αιτήματα.

4.1.1 Χρήσιμα στοιχεία

Όταν ένας χρήστης θέτει ένα αίτημα στη μηχανή αναζήτησης, ξεκινά ένα καινούργιο κομμάτι αιτήματος. Αν ο χρήστης ξαναθέσει το ίδιο αίτημα, ένα ξεχωριστό κομμάτι αιτήματος ξεκινά. Εκεί υπάρχει μια ακολουθία από URLs που συμβολίζονται με U = {u1,…,uM}, όπου το μικρότερο στοιχείο του συνόλου αυτού αναπαριστά την υψηλότερη βαθμολογία σε κατάταξη, δηλαδή πιο κοντά στην κορυφή. Για κανονικά αποτελέσματα αναζήτησης το M συνήθως παίρνει την τιμή 10. Κάθε εμφάνιση ενός URL αντιστοιχίζεται με μία λίστα από τιμές χαρακτηριστικών, όπως είναι η διεύθυνση IP του χρήστη, η τοπική ώρα και η κατηγορία του URL.

4.1.2 Γενικό μοντέλο κλικαρίσματος (General Click Model)

Πρόκειται για μια φωλιασμένη δομή. Το εξωτερικό μοντέλο στη δομή αυτή είναι ένα δίκτυο Bayes στο οποίο υποθέτουμε ότι οι χρήστες σαρώνουν URLs από την κορυφή προς τα κάτω. Οι πιθανότητες μετάβασης του δικτύου αυτού ελέγχονται από ένα εσωτερικό μοντέλο. Πιο συγκεκριμένα, κάθε ξεχωριστή πιθανότητα ορίζεται από το άθροισμα των παραμέτρων, κάθε ένα από το οποίο αντιστοιχεί σε μία μοναδική τιμή χαρακτηριστικού.

4.1.3 Το εξωτερικό μοντέλο

Ξεκινάμε από 1 μέχρι M, οπού το M είναι ο συνολικός αριθμός των URL στη συγκεκριμένη σελίδα. Ορίζουμε δύο δυαδικές τυχαίες μεταβλητές την Ci και την Ei, οι οποίες δείχνουν πότε ο χρήστης κλικάρει ή εξετάζει το URL στην i-στη θέση. Στη συνέχεια, δεσμεύουμε τρεις συνεχείς τυχαίες μεταβλητές σε κάθε θέση, την Ai, Bi και Ri. Η συνεχής συμπεριφορά των μεταβλητών αυτών θα βοηθήσει το μοντέλο να χειριστεί όχι μόνο τη θέση, αλλά και το είδος. Υποθέτουμε ότι ο χρήστης εξετάζει τα URLs από τη θέση i=1 έως i=M. Αφού εξετάσουμε το URL ui (Ei = 1), ο χρήστης επιλέγει να το κλικάρει σύμφωνα με το Ri. Το γεγονός του κλικαρίσματος θα συμβεί αν και μόνο αν το $Ri > 0$. Σε κάθε άλλη περίπτωση, ο χρήστης θα συνεχίσει να εξετάζει το επόμενο URL δηλαδή το u_{i+1} με κάποια πιθανότητα: αν

το ui έχει κλικαριστεί, δηλαδή το Ci=1, ο χρήστης θα εξετάσει το u_{i+1} αν και μόνο αν το Ai $>$ 0, ενώ αν το ui δεν έχει κλικαριστεί, δηλαδή το Ci=0, ο χρήστης θα εξετάσει το URL u_{i+1} και μόνο αν το Bi $>$ 0. Οι ακόλουθες εξισώσεις περιγράφουν το μοντέλο:

$P(E_1) = 1$ $\quad\quad$ $P(E_{i+1} = 1 | E_i = 1, Ci = 1, Ai) = I(Ai > 0)$

$P(E_{i+1} = 1 | E_i = 0) = 0$ $\quad\quad$ $P(Ci=1 | Ei=1, Ri) = I(Ri>0)$

$P(E_{i+1} = 1 | E_i = 1, Ci = 0, Bi) = I(Bi) > 0$

Παρατηρήσεις: το I(.) είναι η χαρακτηριστική συνάρτηση, και ορίζουμε το Φ = {Ai,Bi,Ri|i = 1 …M}.

4.1.4 Το εσωτερικό μοντέλο

Όταν ξεκινάει μια καινούργια δομή αιτημάτων με το αίτημα q και τα urls U = {u1,…,uM}, τα χαρακτηριστικά που η μηχανή αναζήτησης κρατά, απέχουν από το url και τα αιτήματα. Χωρίζουμε τα χαρακτηριστικά αυτά σε δύο κατηγορίες: α) τα χαρακτηριστικά που αφορούν το χρήστη, όπως είναι το αίτημα, η τοποθεσία, το είδος του browser, η τοπική ώρα, η διεύθυνση IP, το μέγεθος του αιτήματος κλπ. Συμβολίζουμε τις τιμές αυτές σαν $f_1^{user},\ldots,f_s^{user}$. β) Τα χαρακτηριστικά που αφορούν τα url, δηλαδή το url, η θέση εμφάνισης(=i), η τάξη του url, η λέξη που ταιριάζει, το μήκος του url κλπ. Για ένα συγκεκριμένο url ui, ορίζουμε τις εξής τιμές χαρακτηριστικών σαν $f_{i,1}^{url},\ldots, f_{i,t}^{url}$.

4.1.5 Ο αλγόριθμος

Έχοντας τη δομή ενός δικτύου Bayes με κρυφές μεταβλητές, η μέθοδος του Expectation Propagation (EP) παίρνει τις τιμές που έχουν παρατηρηθεί σαν είσοδο και είναι ικανός να υπολογίσει το inference οποιασδήποτε μεταβλητής. Η μέθοδος αυτή βρίσκει προσεγγίσεις σε μια κατανομή πιθανότητας. Παρακάτω παρουσιάζεται ο αλγόριθμος:

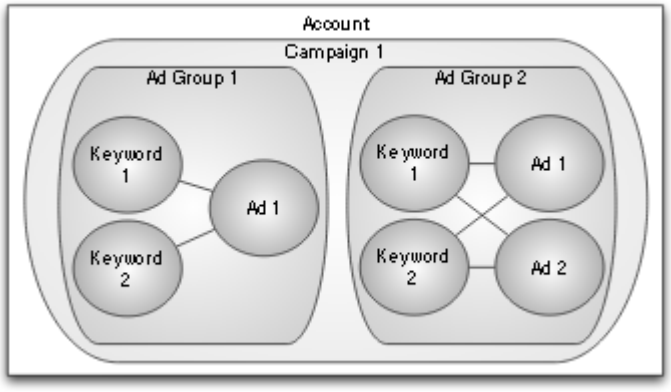
Algorithm: The General Click Model

1. Initiate $\Theta = \{\theta_f^A, \theta_f^B, \theta_f^R | \forall f\}$ and let each parameter in Θ satisfy a prior $N(0, 1/(s+t))$.
2. Construct a Bayesian inference calculator G using *Expectation Propagation*.
3. For each session s
4. $\quad M \leftarrow$ number of urls in s
5. $\quad$ Obtain the attribute values
$$F = \{f_1^{user}, \ldots f_s^{user}\} \cup \{f_{i,1}^{url}, \ldots f_{i,t}^{url}\}_{i=1}^{M}$$
6. $\quad$ Input $\{\theta_f^A, \theta_f^B, \theta_f^R | f \in F\} \subset \Theta$ to G as the prior Gaussian distributions.
7. $\quad$ Input the user's clicks to G as observations.
8. $\quad$ Execute G to measure the posterior distributions for $\{\theta_f^A, \theta_f^B, \theta_f^R | f \in F\}$, and update them in Θ
9. End For

Αναθέτουμε σε κάθε παράμετρο στο σύνολο Θ μια γκαουσιανή κατανομή και ανανεώνουμε την κατανομή αυτή ως εξής: για κάθε κύκλο αιτημάτων, οι υπολογισμένες μεταγενέστερες κατανομές θα χρησιμοποιηθούν σαν τις προγενέστερες κατανομές για τον επόμενο κύκλο αιτημάτων. Στην αρχή του αλγορίθμου, υποθέτουμε ότι όλοι οι παράμετροι στο Θ ικανοποιούν μια γκαουσιανή κατανομή την N(0,1/(s+t)), για όλες τις ξεχωριστές τιμές της συνάρτησης f. Υποθέτουμε επίσης, ότι το εμφωλιασμένο δίκτυο Bayes έχει κατασκευαστεί και το αποτέλεσμα G έχει υπολογιστεί. Θα δουλέψουμε τους κύκλους αιτημάτων έναν προς έναν. Για κάθε καινούργιο κύκλο αιτημάτων που έρχεται, παίρνουμε τη λίστα των τιμών των χαρακτηριστικών $F = \{f_1^{user},\ldots,f_s^{user}\} \cup \{f_{i,1}^{url},\ldots,f_{i,t}^{url}\}_{i=1}^{M}$ και ξαναβρίσκουμε τις αντίστοιχες παραμέτρους στο Θ σαν προγενέστερες κατανομές στο G μαζί με τις σημαίες (flags) για κλικάρισμα ή όχι. Το G θα υπολογίσει τις μεταγενέστερες γκαουσιανές κατανομές θ_f^A, θ_f^B και θ_f^R για κάθε σχετική τιμή χαρακτηριστικού $f \in F$.

4.2 Ιεραρχικό μοντέλο για υπολογισμό τιμών κατά τη διαδικασία του sponsored search

Στο σημείο αυτό, θα μπούμε στη διαδικασία να θυμηθούμε κάποια βασικά στοιχεία για το sponsored search. Καμπάνιες διαφημίσεων, οι οποίες απαρτίζονται από πολλές ομάδες διαφημίσεων (ad groups), είναι το κύριο μοτίβο που εξετάζουμε. Κάθε τέτοια ομάδα διαφημίσεων περιέχει διάφορους όρους (terms) ή λέξεις κλειδιά, για παράδειγμα «αθλητικά παπούτσια», κλπ. Ένα κείμενο διαφημιστικό (creative) αντιστοιχίζεται με ένα τέτοιο σύνολο διαφημίσεων και αποτελείται από έναν τίτλο, μια περιγραφή και ένα URL. Ο τίτλος αποτελείται τυπικά από 2-3 λέξεις και η περιγραφή έχει περίπου 10-15 λέξεις. Ένας διαφημιστής μπορεί στην πραγματικότητα να τοποθετήσει διάφορες περιγραφές σε μια ομάδα διαφημίσεων. Το επόμενο σχήμα δείχνει την ιεραρχική δομή ενός λογαριασμού sponsored search.



Σχήμα 6 : Ιεραρχική δομή ενός λογαριασμού sponsored search [Πηγή: από Βιβλ. Κεφ., αναφορά 2]

Ένας διαφημιστής μπορεί να επιλέξει να χρησιμοποιήσει κανονική (standard) ή προηγμένη (advanced) αντιστοίχηση για τις λέξεις κλειδιά μιας ομάδας διαφημίσεων.

Για παράδειγμα, για κανονική αντιστοίχηση για τη λέξη κλειδί «αθλητικά παπούτσια», θα έχει ως αποτέλεσμα το αντίστοιχο κείμενο περιγραφής (creative) να εμφανίζεται μόνο για αυτό το συγκεκριμένο αίτημα του χρήστη. Όμως, αν υποστηρίζεται προηγμένη αντιστοίχιση, η μηχανή αναζήτησης μπορεί να δείξει την ίδια διαφήμιση για παρεμφερή αιτήματα χρήστη, όπως π.χ. «παπούτσια για τρέξιμο » ή «παπούτσια ορειβασίας».

Θα παρουσιαστεί μια μέθοδος για την πρόβλεψη της αξίας των λέξεων κλειδιών σε μια διαδικασία δύο βημάτων. Πρώτα υπολογίζουμε τις τιμές σε υψηλές σε αξία λέξεις κλειδιά βασιζόμενοι στις προσφορές των διαφημιστών. Στη συνέχεια, θα ταιριάξουμε ένα ιεραρχικό μοντέλο στους μεγαλύτερους από αυτούς τους υπολογισμούς, σχεδιάζοντας τα δημογραφικά καθώς και τα χαρακτηριστικά κειμένου των λέξεων, εκμεταλλευόμενοι την ιεραρχική δομή των λογαριασμών του sponsored search. Έτσι, θα παρουσιαστεί ένα γραμμικό μοντέλο, προκειμένου να βγει σαν συμπέρασμα η αξία κάθε κλικαρίσματος στους όρους που χρησιμοποιούνται για αναζήτηση. Η επιλογή του μοντέλου βασίζεται στην υπόθεση ότι οι λέξεις εκτιμώνται για κάποια χαρακτηριστικά τους, όπως είναι το δημογραφικό προφίλ των χρηστών που προσελκύουν. Το μοντέλο αναπτύσσεται σε δύο στάδια. Στο πρώτο στάδιο, εφαρμόζονται υπάρχουσες τεχνικές από τον τομέα της μικροοικονομίας για να αποκτηθεί η αξία κλικαρίσματος με δεδομένες τις προσφορές των διαφημιστών. Στη συνέχεια, ταιριάζουμε ένα ιεραρχικό γραμμικό μοντέλο σε αυτές τις τιμές που υπολογίστηκαν προηγουμένως. Το αποτέλεσμα είναι ένα μοντέλο, το οποίο μπορεί να δώσει κάποιο συμπέρασμα για οποιονδήποτε καινούργιο όρο, ή μια συλλογή όρων, όσο είναι διαθέσιμα τα απαραίτητα χαρακτηριστικά.

4.2.1 Το μοντέλο χρησιμότητας

Στο σημείο αυτό, περιγράφεται ένα μοντέλο χρησιμότητας για τους διαφημιστές. Πιο συγκεκριμένα, η μεταβλητή i = 1,…,n κατατάσσει τους όρους μιας ομάδας διαφημίσεων. Όταν γίνεται προσφορά σε έναν όρο i, ο διαφημιστής επιλέγει μια επαρκή ποσότητα κλικαρισμάτων x_i. Η πιθανότητα $p_i(x_i)$ εκφράζει το αναμενόμενο συνολικό κόστος για να αποκτηθούν αυτά τα κλικ. Αυτή η χρησιμότητα για τα κλικ, που αποκτάται από τους όρους της ομάδας των διαφημίσεων, παίρνουμε την εξής μορφή:

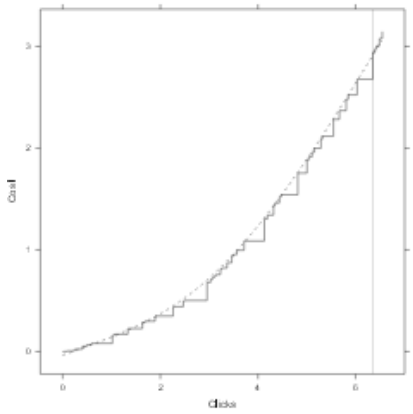
$U(x) = [\sum_{i=1}^{n} v_i {x_i}^{1-\rho}]^{1/1-\rho} - \sum_{i=1}^{n} p_i(x_i)$. Ο πρώτος όρος παίρνει τη γνωστή μορφή της συνάρτησης χρησιμότητας CES (Constant Elasticity of Substitution). Έχει μια παράμετρο v_i για κάθε όρο i, όπου υπό μία έννοια δεσμεύει την αξία των κλικ από κάθε όρο. Επίσης, έχει και μια καθολική παράμετρο ρ, όπου διαμορφώνει την ελαστικότητα ανάμεσα στα κλικ πάνω σε διαφορετικούς όρους. Στο p=0, το στοιχείο είναι προσθετικό, και η χρησιμότητα αποσυνδέεται ανάμεσα στους όρους, ώστε να ληφθούν υπόψη απομονωμένα. Στο $p \to +\infty$ φτάνουμε σε καθαρά στοιχεία. Αν κάποιος διαφημιστής προσφέρει σε υψηλές τιμές σε έναν όρο, ώστε να πετυχαίνει το μεγαλύτερο slot σε κάθε δημοπρασία, τότε έχει αποκτηθεί η μεγαλύτερη δυνατή ποσότητα κλικαρισμάτων από αυτόν τον όρο. Αφήνοντας το X_i να δηλώνει το μέγιστο, το σύνολο κατανάλωσης του διαφημιστή διαμορφώνεται ως εξής: $x_i \leq X_i$, (i= 1,…,n). Η παραπάνω εξίσωση μαζί με αυτόν τον περιορισμό που παρουσιάστηκε τώρα αποτελούν το πρόβλημα ζήτησης του διαφημιστή.

Πρόταση 1

Υποθέτουμε ότι κάθε συνάρτηση τιμής p_i ότι είναι διαφορίσιμη. Τότε σε μια βέλτιστη λύση $x^* > 0$ του προβλήματος ζήτησης του διαφημιστή έχουμε: logvi ≥ logpi'(xi*) + plogxi* + pC, για i = 1,…,n, με ισότητα αν το $xi^* < Xi$. Εδώ το C είναι μια σταθερά ανεξάρτητη από τους όρους. Ο περιορισμός ότι το $x^* > 0$, η βέλτιστη επιλογή κλικαρισμάτων είναι απόλυτα θετική είναι χωρίς βλάβη της γενικότητας για τους σκοπούς μας, επειδή σε εμπειρική ανάλυση θα λάβουμε υπόψη μας όρους στους οποίους ο διαφημιστής συναντά την αποθεματική τιμή. Επίσης, να σημειωθεί ότι η πιο πάνω ανισότητα είναι μια ισότητα για όλους τους όρους, εκτός από αυτούς όπου ο διαφημιστής φτάνει τη μέγιστη ποσότητα κλικαρισμάτων. Όταν το p=0, ανακτούμε την γνωστή μας συνθήκη ότι η οριακή αξία ισούται με την οριακή τιμή σαν βέλτιστη επιλογή των κλικ. Έως εδώ έχουμε μοντελοποιήσει την χρησιμότητα ενός διαφημιστή σε επίπεδο ομάδας διαφημίσεων. Για μεγαλύτερα επίπεδα, υποθέτουμε ότι η χρησιμότητα είναι προσθετική απέναντι στις ομάδες διαφημίσεων, ώστε αυτές να λαμβάνονται υπόψη μεμονωμένα. Τέλος, η μεθοδολογία υπολογισμού της αξίας είναι ακόμα σχετική, εάν ένας διαφημιστής δεν ενδιαφέρεται για τα χρήματα. Με περιορισμό για τα χρήματα η πιθανότητα pi(xi*), που έχει χρησιμοποιηθεί σε παραπάνω εξίσωση, χρειάζεται να αντικατασταθεί με τον όρο λpi(xi*), όπου το λ ≥ 0 είναι ένας πολλαπλασιαστής.

4.2.2 Εκτίμηση της αξίας (value estimation)

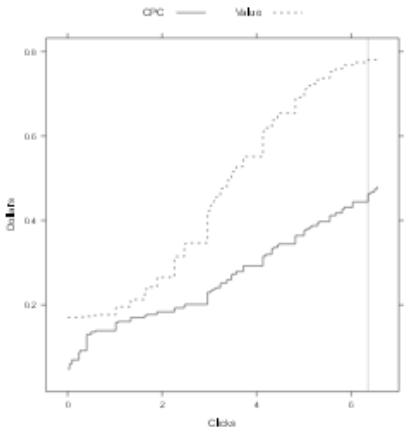
Το $\hat{x}$i(bi) δείχνει τα συνολικά αναμενόμενα κλικ των διαφημιστών, όταν προσφέρουμε bi για τον όρο i, και ομοίως το $\hat{p}$i(bi) δείχνει το συνολικό αναμενόμενο κόστος. Υπογραμμίζουμε ότι αυτά είναι σύνολα στη σχετική χρονική περίοδο και όχι οι τιμές του CTR (click-through rate) και CPC (cost per click). Δουλεύουμε ακόμη για τη χρονική περίοδο ενός μήνα. Χρησιμοποιούμε πληροφορία από πραγματική αναζήτηση που συνέβη έναν συγκεκριμένο μήνα παρατήρησης. Υποθέτουμε ότι ο διαφημιστής παίρνει την απόφαση να τοποθετήσει μια μοναδική προσφορά για κάθε όρο για ολόκληρο τον μήνα. Η θέση του διαφημιστή σε μια λέξη κλειδί μπορεί να ποικίλλει με το χρόνο, επειδή οι αντίπαλοι αλλάζουν και η ποιότητα ανανεώνεται συστηματικά. Για κάθε στιγμιότυπο δημοπρασίας διατρέχουμε τις προσφορές που ο διαφημιστής θα μπορούσε υποθετικά να εφαρμόσει μέσα σε κάποια σχετικά πλαίσια, όπως [bmin, bmax] και προσομοιώνουμε τις τιμές για το CTR και CPC που θα παίρναμε για αυτήν την προσφορά. Μπορούμε να προσομοιώσουμε τις αλλαγές στη θέση, επειδή διαθέτουμε τις προσφορές του αντιπάλου και τη βαθμολογία της ποιότητας και έτσι παίρνουμε το προβλεπόμενο CTR από τους υπολογισμούς της μηχανής αναζήτησης. Ορίζουμε μια αντίστροφη συνάρτηση ως εξής: $\hat{x}i^{-1}(x_i)$ = inf{bi: x(bi)≥xi}. Το αναμενόμενο συνολικό κόστος σαν συνάρτηση κλικαρισμάτων, το οποίο είναι ένα από τα δύο κλειδιά, ανακατασκευάζεται στη συνέχεια από το $\hat{x}i$ και το $\hat{p}i$ και έχουμε pi(xi) = $\hat{p}i(\hat{x}i^{-1}(x_i))$. Αναφερόμαστε σε αυτήν την συνάρτηση σαν μια εκτιμόμενη καμπύλη προσφοράς. Κληρονομεί τις ιδιότητες του $\hat{x}i$ και $\hat{p}i$ και είναι μια βηματική συνάρτηση κλικαρισμάτων, και για υψηλούς όρους τείνει να γίνει συνεχής. Τελικά, προσεγγίζουμε το pi στα σημεία που αλλάζει τιμή, με χρήση κυβικής παρεμβολής, και καταλήγει σε μια καμπύλη με ομαλές παραγώγους.



Σχήμα 7 : Καμπύλη προσφοράς [Πηγή: από Βιβλ. Κεφ., αναφορά 2]

Από την καμπύλη προσφοράς μπορούμε να βρούμε το οριακό κόστος, το οποίο αντιστοιχεί στις εκτιμόμενες τιμές, όπως επίσης και το μέσο κόστος για κάθε κλικ.

Το σχήμα 8 παρουσιάζει τις εκτιμήσεις που προκύπτουν από την καμπύλη προσφοράς.



Σχήμα 8 : Καμπύλη εκτίμησης τιμών [Πηγή: από Βιβλ. Κεφ., αναφορά 2]

Με δεδομένους τους κανόνες της δημοπρασίας με sponsored search, αναμένουμε το μέσο CPC να αυξάνεται με τα κλικαρίσματα. Είναι ένα βασικό αποτέλεσμα της μικροοικονομίας, όπου η αύξηση του μέσου κόστους υπονοεί οριακό κόστος να είναι πάνω από το μέσο κόστος. Το μοντέλο αυτό είναι κατά τον διαφημιστή κοντά στο ιδανικό, αφού βασίζεται σε πληροφορία από την ίδια χρονική περίοδο κατά τη διάρκεια της οποίας ο διαφημιστής προσφέρει, και δίνει μια προσφορά και πληροφορία από ακολουθία κλικαρισμάτων, καθώς επίσης και τους υπολογισμούς των τιμών του CTR από τη μηχανή αναζήτησης. Στην πράξη, οι μετρήσεις μας στην επιλεγμένη ποσότητα των κλικ για τη δεδομένη χρονική περίοδο είναι απλά $x_i = \widehat{xi}(\overline{bi})$, όπου το $\overline{bi}$ είναι η προσφορά του διαφημιστή ενάντια στις προσφορές στον όρο i.

4.2.3 Διαγνωστικός έλεγχος (Diagnostic checks)

Μη αρνητικές τιμές: Ίσως είναι ο πιο απλός τρόπος για έναν έλεγχο οι τιμές να είναι μη αρνητικές, αλλιώς θα είναι χωρίς νόημα να προσφέρουμε σε όρους που δεν έχουν αξία. Η κυβική παρεμβολή και οι παράγωγοι μπορούν να μετατραπούν σε αρνητικές τιμές.

Υπερπροσφορά (overbids): Αναφερόμαστε σε ένα στιγμιότυπο, όπου η προσφορά υπερβαίνει την εκτιμώμενη τιμή ως υπερπροσφορά (overbid). Οι υπερπροσφορές αντιστοιχούν σε προσφορές που πετυχαίνουν την μεγαλύτερη ποσότητα κλικαρισμάτων σε έναν όρο δεν είναι κάτι αξιοπρόσεχτο και αυτό γιατί, σε αυτήν την περίπτωση, οι τιμές που έχουν εκτιμηθεί είναι χαμηλά όρια πάνω σε πραγματικές τιμές.

Ατομική λογική (individual rationality): Αυτό που μας απασχολεί περισσότερο είναι οι όροι που η εκτιμόμενη αξία πέφτει κάτω από την τιμή CPC, με αποτέλεσμα να οδηγούμαστε σε αρνητικό κέρδος για κάθε κλικ και έτσι να έχουμε παραβίαση της ατομικής λογικής.

Ανάλυση ευαισθησίας (sensitivity analysis): Κάνουμε αυτού του είδους την ανάλυση για να εξετάσουμε πώς η αξία και οι υπολογισμοί στα κλικ του χρήστη συμπεριφέρονται στις προσφορές του διαφημιστή. Η ευρωστία στις προσφορές είναι κάτι σημαντικό καθώς αυτές οι παρατηρήσεις είναι σημαντικές είσοδοι για το ιεραρχικό μοντέλο που θα παρουσιαστεί στη συνέχεια.

4.2.4 Το ιεραρχικό μοντέλο

Η υπόθεση πίσω από το μοντέλο αυτό είναι ότι η αξία του διαφημιστή στα κλικαρίσματα από μια λέξη κλειδί μπορεί να αποσυντεθεί σύμφωνα με τις ιδιότητες των τιμών. Στο σημείο αυτό, θα εισάγουμε ένα ιεραρχικό γραμμικό μοντέλο για να προσδιορίσουμε την σχέση ανάμεσα στην προσφορά ή την τιμή σε κάποιο μετρήσιμο χαρακτηριστικό των λέξεων κλειδιά. Κατατάσσουμε αρχικά τους όρους σε μια καμπάνια ως προς i, και τις ομάδες διαφημίσεων κατά j. Χρησιμοποιούμε τον συμβολισμό j[i] για να δείξουμε την ομάδα διαφημίσεων στην οποία ανήκει κάποιος όρος i. Για κάθε όρο i δείχνουμε το διάνυσμα των χαρακτηριστικών και το συμβολίζουμε με xi. Το μοντέλο των όρων λοιπόν παίρνει την εξής μορφή: $y_i = a_{j[i]} + \beta' x_i + \varepsilon_i$, όπου το $\varepsilon_i \sim N(0,\sigma^2_y)$. Για εδώ έχουμε ότι $y_i = \log b_i$, που είναι ο φυσικός λογάριθμος της προσφοράς για τον όρο i. Δείχνουμε το διάνυσμα του επιπέδου της ομάδας διαφημίσεων που περιέχει τα χαρακτηριστικά με u_j για μια ομάδα διαφημίσεων j. Την ομάδα διαφημίσεων ακολουθεί το εξής μοντέλο: $a_j = v + \delta' u_j + \varepsilon_j$, όπου το $\varepsilon_i \sim N(0,\sigma^2_\alpha)$. Προκειμένου να αναπτυχθεί το μοντέλο για την αξία και όχι για τις προσφορές, με χρήση της εξίσωσης που παρουσιάστηκε παραπάνω και συγκεκριμένα στην πρόταση 1, παίρνουμε ότι yi = logpi'(xi*), εισάγοντας το logxi* σαν μια επιπλέον πρόβλεψη στο δεξί μέλος. Η παράμετρος p τότε θα αντιστοιχεί στον συντελεστή που αφορά τα κλικ. Επίσης, ο σταθερός όρος στην εξίσωση αυτή ενσωματώνεται στο σταθερό όρο της οπισθοδρόμησης. Το επόμενο βήμα είναι να εισαχθούν κάποια προγνωστικά που θα αφορούν την προσφορά και την αξία. Κάποια από αυτά που θα παρουσιαστούν παρακάτω χρησιμοποιούνται και σε επίπεδο όρων. Ενδεικτικά αναφερόμαστε στην ηλικία, στο εισόδημα, στο φύλο, στο ακριβές ταίριασμα, στο μέγεθος της ουράς κλπ. Περισσότερες λεπτομέρειες περιγράφονται στη σχετική δημοσίευση [Πηγή: από Βιβλ. Κεφ., αναφορά 2].

4.3 Πρόβλεψη εμφάνισης διαφημίσεων κατά τη διαδικασία του sponsored search (Ad impression forecasting for sponsored search)

Στο σημείο αυτό, αναπτύσσεται ένα παραγωγικό μοντέλο. Γίνεται χρήση του δικτύου Bayes για να καταγραφούν οι εξαρτήσεις ανάμεσα στα χαρακτηριστικά που αφορούν το αίτημα του χρήστη και τους ανταγωνιστές σε μια διαφημιστική δημοπρασία. Επίσης, χρησιμοποιείται ένα γραμμικό δυναμικό μοντέλο προκειμένου να φανεί η μεταβλητότητα των αιτημάτων των χρηστών. Η προσέγγιση αυτή γίνεται στο πλαίσιο της διαδικασίας MapReduce, η οποία παρουσιάζεται εκτενώς στη βιβλιογραφία. [Πηγή: από Βιβλ. Κεφ., αναφορά 6]

4.3.1 Εισαγωγικά και σημειογραφία

Στο σημείο αυτό, θα εισάγουμε κάποιες έννοιες και συμβολισμούς που θα χρησιμοποιήσουμε στη συνέχεια. Αναφερόμαστε λοιπόν σε μια διαφήμιση στο σύστημα του sponsored search σαν μια καταχώρηση και την συμβολίζουμε με L. Κάποια χαρακτηριστικά του αιτήματος του χρήστη συμβολίζονται με Q. Κάποια από αυτά είναι η τοποθεσία, η κατηγορία και η ώρα που τέθηκε το ερώτημα. Ο συμβολισμός bid(L) εκφράζει την αξία προσφοράς της καταχώρησης L, και το pclick(L,Q) εκφράζει την εκτιμόμενη πιθανότητα του κλικαρίσματος στην καταχώρηση L, όταν ο χρήστης θέτει ένα αίτημα με χαρακτηριστικά Q. Η μηχανή εξυπηρέτησης των διαφημίσεων διεξάγει μια δημοπρασία για κάθε αίτημα χρήστη. Για κάθε τέτοια φράση λοιπόν, οι διαφημιστές ανταγωνίζονται με τις προσφορές και το αποτέλεσμα για κάθε διαφήμιση L δίνεται από τον τύπο : score(L,Q) = bid(L) x pclick(L,Q). Το pclick(L,Q) υπολογίζεται χρησιμοποιώντας έναν αλγόριθμο που κάνει χρήση διάφορων χαρακτηριστικών από το αίτημα Q, αλλά και από την καταχώρηση L και τον αντίστοιχο διαφημιστή. Μετά τον υπολογισμό του σκορ για κάθε καταχώρηση, δοσμένου ενός αιτήματος, τα αποτελέσματα ταξινομούνται στις k καταχωρήσεις με το υψηλότερο σκορ και επιλέγονται αυτές. Το k είναι μια παράμετρος της μηχανής αναζήτησης και μπορεί να διαφέρει ανάμεσα στις διάφορες δημοπρασίες. Μπορεί να υπάρχουν και επιπλέον κριτήρια προκειμένου να επεκταθεί αυτή η λίστα των διαφημίσεων. Αν κάποια διαφήμιση κλικαριστεί από κάποιον χρήστη σε μια δημοπρασία Q, τότε ο διαφημιστής δημιουργεί μια διαδικασία πληρωμής στην μηχανή αναζήτησης. Η πληρωμή αυτή υπολογίζεται από τη μέθοδο GSP: payment(Li,Q) = score(L_{i+1},Q) / pclick(L_i,Q), όπου το L_{i+1} είναι η καταχώρηση μετά την L_i στην ταξινομημένη λίστα των σκορ.

4.3.2 Μεθοδολογία

Αρχικά, θα περιγράψουμε το πρόβλημα πρόβλεψης εμφάνισης των διαφημίσεων και στη συνέχεια θα προτείνουμε μια μέθοδο για το πρόβλημα αυτό.

Για έναν διαφημιστή A, τη διαφήμισή του L και την αξία προσφοράς bid(L), σκοπός της πρόβλεψης της εμφάνισης των διαφημίσεων είναι για να προβλέψουμε τον αριθμό των εμφανίσεων μιας διαφήμισης L σε ένα σταθερό χρονικό διάστημα στο μέλλον. Για απλότητα, υποθέτουμε ότι τα διάστημα αυτό είναι μια βδομάδα. Η πληροφορία που χρησιμοποιείται για τη μέθοδό μας προέρχεται από αρχεία δημοπρασιών που περιέχουν πληροφορία από όλες τις δημοπρασίες του πρόσφατου παρελθόντος, όπως τα

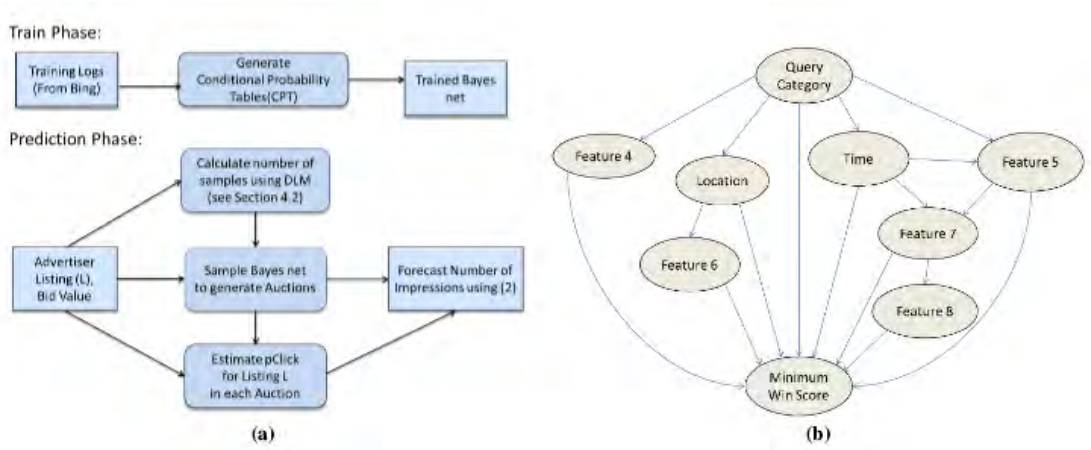
χαρακτηριστικά των αιτημάτων και τα σκορ των βαθμολογιών. Περιγράφουμε κάποια βασικά χαρακτηριστικά της δημοπρασίας :

α) Το αίτημα του χρήστη (query): Οι μηχανές αναζήτησης βγάζουν τυπικά κάποια χαρακτηριστικά από το αίτημα του χρήστη. Για παράδειγμα, τα χαρακτηριστικά αυτά μπορεί να είναι η τοποθεσία από όπου προέρχεται το αίτημα, η ώρα που στάλθηκε το αίτημα ή η κατηγορία. Αυτά τα χαρακτηριστικά είναι σημαντικά, διότι σε μια δημοπρασία οι ανταγωνιστές βαθμολογούν τις καταχωρήσεις L και αυτά παίζουν σημαντικό ρόλο στο αποτέλεσμα του σκορ. Για παράδειγμα, υποθέτουμε ότι η καταχώρηση L έχει κάποια κλικ από χρήστες στην Νέα Υόρκη. Έτσι, αν το αίτημα του χρήστη προέρχεται από τη Νέα Υόρκη, τότε το pclick του L θα είναι υψηλό, εξ ‘αιτίας του υψηλού σκορ για την κατάταξη L που υπάρχει σε αυτές τις δημοπρασίες.

β) Το σκορ του ανταγωνιστή: μοντελοποιούμε το σκορ των «τυπικών» ανταγωνιστών μιας δοσμένης κατάταξης L. Για να προβλέψουμε τον αριθμό των εμφανίσεων, χρειάζεται να προβλέψουμε πότε μια διαφήμιση θα «κερδίσει» σε μια συγκεκριμένη δημοπρασία, δηλαδή να προβλεφθεί το ελάχιστο σκορ που χρειάζεται για να κερδηθεί μια δημοπρασία. Έτσι, μοντελοποιούμε αυτό το ελάχιστο σκορ, το οποίο λαμβάνει υπόψη του διάφορους παράγοντες, όπως είναι οι προσφορές και τα pclicks των ανταγωνιστών, τον αριθμό των νικητών, καθώς και άλλα κριτήρια.

Υπολογίζουμε το σκορ πολλαπλασιάζοντας την τιμή προσφοράς και την τιμή του pclick. Όμως, για τον υπολογισμό του pclick συνήθως εμπλέκεται και ο χρόνος. Έτσι, σε διαφορετικές στιγμές το pclick μιας διαφήμισης σε μια δημοπρασία για το ίδιο αίτημα μπορεί να διαφέρει. Στη συνέχεια, περιγράφεται ένα μοντέλο παραγωγής για να μοντελοποιήσουμε τις δημοπρασίες στις οποίες συμμετέχει μια διαφήμιση L. Στο σημείο αυτό χρησιμοποιούμε το δίκτυο Bayes, όπου κάθε κόμβος του δικτύου αντιστοιχεί είτε σε κάποιο χαρακτηριστικό του αιτήματος του χρήστη, είτε στο ελάχιστο σκορ που χρειάζεται για να κερδηθεί η δημοπρασία. Πιο συγκεκριμένα, το δίκτυο Bayes αναπαρίσταται από κατευθυνόμενο ακυκλικό γράφημα, όπου κάθε κόμβος του είναι μια τυχαία μεταβλητή ή ένα χαρακτηριστικό του αιτήματος στο σύστημα μας. Οι ακμές ανάμεσα σε διαφορετικούς κόμβους αιχμαλωτίζουν τις εξαρτήσεις ανάμεσα σε διαφορετικούς κόμβους. Ειδικά, ικανοποιείται την μαρκοβιανή ιδιότητα, ότι δηλαδή δεδομένων των γονιών, ένας κόμβος είναι ανεξάρτητος από όλους τους άλλους κόμβους που δεν είναι απόγονοί του σε αυτό το δίκτυο Bayes. Έτσι, Pr(vi/Fi) = Pr(vi/Gi), όπου το Fi = {vj/vi είναι παιδί του vj } και το Gi = {vk/vk δεν είναι απόγονος του vi}. Τώρα, κάθε μέλος της κοινής διανομής μπορεί να υπολογιστεί χρησιμοποιώντας τα παραπάνω. Έτσι, για να τεστάρουμε ένα δίκτυο Bayes, απλά υπολογίζουμε την πιθανότητα Pr(vi/Fi) σε κάθε κόμβο vi.

Στο παρακάτω σχήμα φαίνεται ένα απλό δίκτυο Bayes που χρησιμοποιούμε. Δεν αποκαλύπτονται χαρακτηριστικά του αιτήματος του χρήστη για λόγους εμπιστοσύνης. Το ελάχιστο σκορ της δημοπρασίας είναι ένας κόμβος παιδί για κάθε κόμβο με χαρακτηριστικά που αφορούν το αίτημα του χρήστη.



Σχήμα 9 : Παράδειγμα δικτύου Bayes [Πηγή: από Βιβλ. Κεφ., αναφορά 3]

Στο σχήμα 9α έχουμε την επισκόπηση της φάσης δοκιμής (train phase) καθώς και της φάσης της πρόβλεψης (prediction phase) για το μοντέλο παραγωγής (Generative Model), το οποίο βασίζεται στην μέθοδο Ad Impression Forecasting Method (GMIF). Στην φάση δοκιμής, η μέθοδος GMIF χρησιμοποιεί αρχεία δοκιμών για να παράγει τους πίνακες πιθανοτήτων (CPT) για κάθε κόμβο του δικτύου Bayes. Κατά τη διάρκεια της φάσης πρόβλεψης, δοσμένου ενός καταλόγου L και μιας τιμής προσφοράς, η μέθοδος GMIF δειγματοληπτεί το δίκτυο Bayes για να παράγει τις δημοπρασίες. Ο αριθμός των δειγμάτων που απαιτούνται, προσδιορίζεται από το στοιχείο DLM. Η τιμή του pclick για τον κατάλογο L για κάθε δημοπρασία που παράγεται, υπολογίζεται με χρήση του τελευταίου στοιχείου παραγωγής. Με χρήση της παρεχόμενης τιμής προσφοράς pclick, η μέθοδος GMIF υπολογίζει την βαθμολογία του καταλόγου L κάθε δημοπρασίας και προβλέπει τον αριθμό των εμφανίσεων χρησιμοποιώντας το στοιχείο (2) του σχήματος.

Στο σχήμα 9b είναι ένα παράδειγμα δικτύου Bayes για να μοντελοποιήσουμε τις δημοπρασίες κατά τη διαδικασία του Sponsored Search.

Όπως επισημάνθηκε και νωρίτερα, για το δίκτυο Bayes, υπολογίζουμε την ποσότητα CPT για κάθε κόμβο, χρησιμοποιώντας υλικό που υπάρχει στο αρχείο. Εξετάζουμε ένα δίκτυο Bayes για κάθε λέξη κλειδί. Στη συνέχεια, δοσμένης μιας διαφήμισης L, χρησιμοποιούμε το δίκτυο Bayes, το οποίο είναι σχεδιασμένο για κάθε λέξη προσφοράς να παράγει μια δημοπρασία. Ο αριθμός των δημοπρασιών αυτών εκτιμάται από το γραμμικό δυναμικό μοντέλο (Dynamic Linear Model-DLM), το οποίο βασίζεται σε μέθοδο που θα παρουσιάσουμε στη συνέχεια. Ακολούθως, μετά την παραγωγή των δημοπρασιών, τοποθετούμε τη διαφήμιση L σε κάθε δημοπρασία και υπολογίζουμε την pclick για κάθε δημοπρασία, χρησιμοποιώντας την τελευταία παραγωγή μονάδας του pclick. Μετά, αφού έχουμε την τιμή προσφοράς bid(L), υπολογίζουμε το score(L,Q) σε κάθε δημοπρασία και υπολογίζουμε τον συνολικό αριθμό εμφανίσεων, συγκρίνοντας με το ελάχιστο σκορ το οποίο κερδίζει στην δημοπρασία.

Ας πούμε ότι το M_L είναι ο εκτιμόμενος αριθμός δημοπρασιών που παράγονται από την κατάταξη L. Έτσι, τα σύνολα των δημοπρασιών είναι: $A = \{A_1,A_2,\ldots,A_{ML}\}$, όπου $A_i = \{Q_i,MS_i\}$, το Q_i αποτελείται από τα χαρακτηριστικά του αιτήματος του χρήστη για την δημοπρασία A_i και το MS_i είναι το ελάχιστο σκορ που χρειάζεται για να κερδίσει την A_i. Επίσης, ας ορίσουμε το

υπολογισμένο pclick της διαφήμισης L στην δημοπρασία A_i να ισούται με το pclick(L,Q_i), για κάθε $1 \leq i \leq M_L$. Έτσι, ο συνολικός αριθμός των εμφανίσεων δίνεται από τον τύπο: impressions(L) = |{i|bid(L) x pclick(L,Q_i) ≥ MS_i}|. Από το παραπάνω σχήμα φαίνεται μια περίληψη της φάσης πρόβλεψης και της φάσης προετοιμασίας. Να επισημανθεί ότι ο προβλεπόμενος αριθμός εμφανίσεων εξαρτάται από το ML, δηλαδή ο εκτιμόμενος αριθμός δημοπρασιών που η διαφήμιση L θα συμμετέχει στο τεστ της εβδομάδας. Η εκτίμηση για το ML αποτελεί μια πρόβλεψη εξ αιτίας δύο γεγονότων.

Πρώτον, ο αριθμός των ερευνών για ένα αίτημα του χρήστη μπορεί να διαφέρει ανάμεσα σε διαφορετικές χρονικές περιόδους. Αυτό το φαινόμενο είναι ιδιαίτερα κυρίαρχο στις λεγόμενα αιτήματα ουράς (tail queries), τα οποία είναι αρκετά δημοφιλή για μικρές χρονικές περιόδους. Για παράδειγμα, αιτήματα που αφορούν ένα όνομα ταινίας φτάνει σε μέγιστο όριο κυρίως όταν βγαίνει η ταινία και στη συνέχεια πέφτει με μεγάλο ρυθμό καθώς περνούν οι εβδομάδες.

Δεύτερον, μια κατάταξη διαφήμισης L μπορεί και να μην συμμετέχει σε κάθε δημοπρασία για κάθε λέξη κλειδί προσφοράς για διάφορους λόγους. Για παράδειγμα, ένας διαφημιστής μπορεί να θέλει να προσελκύσει ένα συγκεκριμένο κοινό. Επίσης, η μηχανή αναζήτησης για sponsored search μπορεί να θέλει να φιλτράρει μια διαφήμιση, εξ αιτίας του ότι δεν σχετίζεται σε μεγάλο βαθμό με κάποιο συγκεκριμένο αίτημα ή να εισάγει κάποιας μορφής τυχαιότητα για έρευνα.

4.3.3 Υπολογισμός του αριθμού των δημοπρασιών

Στο σημείο αυτό, περιγράφεται η μέθοδος για τον υπολογισμό του αριθμού των δημοπρασιών, στο οποίο η δοσμένη κατάταξη L είναι πιθανόν να συμμετέχει στο τεστ της εβδομάδας. Χειριζόμαστε το πρόβλημα με το να το διασπάσουμε σε δύο προβλήματα. Το πρώτο, είναι να καθοριστεί ο αριθμός των ερευνών που θα πραγματοποιηθούν για τη διαφήμιση L της λέξης κλειδί προσφοράς, και το δεύτερο είναι να προσδιοριστεί η συχνότητα συμμετοχής για το L, δηλαδή, το όριο των δημοπρασιών, στο οποίο το L είναι πιθανόν να συμμετέχει. Όσον αφορά το πρώτο πρόβλημα, η προσέγγιση μας είναι να μοντελοποιήσουμε την ποσότητα των ερευνών για μια λέξη κλειδί, με χρήση της πρώτης τάξης γραμμικό δυναμικό μοντέλο (Dynamic Linear Model-DLM) προκειμένου να προσδιοριστεί το επόμενο σημείο. Μετά από όλα αυτά, θα έχουμε αποκτήσει τον συνολικό αριθμό των δημοπρασιών M_L, όπου το L είναι πιθανό να συμμετέχει. Επίσης, M_L = N x γ_L, όπου το N είναι ο εκτιμόμενος αριθμός ερευνών και το γ είναι η υπολογισμένη συχνότητα συμμετοχής του L.

4.3.4 Το δυναμικό γραμμικό μοντέλο

Στο σημείο αυτό, θα περιγράψουμε τη μέθοδο του αριθμού των ερευνών για μια λέξη κλειδί. Προκειμένου, να σχηματιστεί η χρονική συνέχεια, χωρίζουμε τη γραμμή του χρόνου σε σημεία, κάθε ένα μεγέθους μιας βδομάδας. Στη συνέχεια, με χρήση των αρχείων δεδομένων, υπολογίζουμε τον αριθμό των ερευνών για κάθε λέξη κλειδί και για κάθε εβδομάδα. Το Nt υποδηλώνει τον αριθμό των ερευνών την t-th βδομάδα της λέξης κλειδί προσφοράς, που αντιστοιχεί στη διαφήμιση L. Επίσης, $1 \leq t \leq T$, όπου το T είναι η βδομάδα των δοκιμών για την οποία θέλουμε να προβλέψουμε τον αριθμό των δημοπρασιών.

Δοκιμάζουμε το μοντέλο αυτό χρησιμοποιώντας $\{N_1,\ldots,N_t,\ldots,N_{t-1}\}$ και προβλέποντας τον αριθμό των ερευνών στην T εβδομάδα. Στη συνέχεια, περιγράφουμε την πρώτου επιπέδου μέθοδο για την πρόβλεψη. Η μέθοδος αυτή παρακινείται από την παρατήρηση ότι τα κομμάτια που προκαλούν συμφόρηση είναι μικρά σε ζωή και έτσι η μέθοδος αυτή ταιριάζει σε βραχυπρόθεσμες προβλέψεις. Έτσι λοιπόν, ο αριθμός των ερευνών για μια συγκεκριμένη λέξη κλειδί σε χρόνο t δίνεται από τον τύπο: $N_t = \mu_t + \nu_t \sim N(0,V)$ και $\mu_t = \mu_{t-1} + \omega_t$, $\omega_t \sim N(0,W)$, όπου το μt είναι η εσωτερική κατάσταση της σειράς, V,W > 0 είναι σταθερές και το N(0,V) είναι η γκαουσιανή κατανομή με μέση τιμή 0 και διασπορά V. Υποθέτουμε ότι μ0 ~ N (0, C0), όπου το C0 > 0 είναι μια σταθερά. Στη συνέχεια, με χρήση του μοντέλου αυτού προκύπτουν οι ακόλουθες εξισώσεις:

$(N_t | N_1,\ldots,N_{t-1}) \sim N(m_{t-1}, C_{t-1} + V + W)$,

$Mt = m_{t-1} + \frac{C_{t-1}+W}{C_{t-1}+W+V}(N_{t-1} - m_{t-1})$,

$Ct = \frac{(C_{t-1}+W)V}{C_{t-1}+W+V}$, όπου το m0 = 0. Το Nt είναι μια τυχαία μεταβλητή που αντιστοιχεί στον αριθμό των ερευνών μια λέξης κλειδί.

4.3.5 Υπολογισμός της συχνότητας συμμετοχής

Στο σημείο αυτό, περιγράφουμε τη μέθοδο για να υπολογίζουμε τη συχνότητα συμμετοχής γ_L της δοσμένης διαφήμισης L, η οποία ορίζεται σαν τη συχνότητα του αριθμού των δημοπρασιών, στην οποία το L συμμετέχει στην T βδομάδα προς τον συνολικό αριθμό των δημοπρασιών για την λέξη κλειδί προσφοράς της διαφήμισης L. Να σημειωθεί ότι, σε πραγματικά συστήματα μια διαφήμιση δεν συμμετέχει σε όλες τις δημοπρασίες εξ αιτίας διαφόρων λόγων, όπως είναι περιορισμοί στα χρήματα που θα δαπανηθούν, το φιλτράρισμα από το σύστημα sponsored search κλπ. Για παράδειγμα, αν το μπατζετ για μια διαφήμιση έχει τελειώσει, τότε η συγκεκριμένη διαφήμιση δε θα μπορέσει να συμμετέχει σε μελλοντικές δημοπρασίες για τη λέξη κλειδί αυτή. Παρόμοια, οι διαφημιστές μπορούν να παρέχουν κάποιους περιορισμούς προκειμένου να προσελκύσουν μια συγκεκριμένη ομάδα χρηστών. Συνεπώς, στην πράξη, η συχνότητα συμμετοχής τείνει να είναι μικρή για κάποιες διαφημίσεις. Έτσι, ο υπολογισμός της συχνότητας συμμετοχής γL είναι μια σημαντική παράμετρος για τη μέθοδό μας. Να σημειωθεί πως ένας άλλος τρόπος υπολογισμού του γ_L είναι μέσα από τη χρήση μεθόδων πρόβλεψης, με βάση την πάροδο του χρόνου. Παρόλο που η χρονική διάρκεια στις περισσότερες διαφημίσεις είναι γύρω στις δύο εβδομάδες, δε μπορούμε να εφαρμόσουμε το δυναμικό γραμμικό μοντέλο σωστά για το πρόβλημα αυτό. Έτσι, κάνουμε την απλή υπόθεση ότι το γL παραμένει σταθερό καθώς περνάει ο χρόνος. Με τον τρόπο αυτό υπολογίζεται στη συνέχεια το γL κάνοντας χρήση των δεδομένων που υπάρχουν στο αρχείο. Αυτό που γίνεται είναι ο υπολογισμός του συνολικού αριθμού των νικών της διαφήμισης L και να διαιρεθεί από τον συνολικό αριθμό των δημοπρασιών που συμμετέχει η διαφήμιση L στη εβδομάδα των πειραμάτων. Η μέθοδος αυτή εφαρμόζεται σε ήδη υπάρχουσες διαφημίσεις. Αν οι διαφημίσεις δεν έχουν εμφανιστεί στα αρχεία τα οποία χρησιμοποιούμε για να εφαρμόσουμε τη μέθοδο μας, τότε μιλάμε για καινούργιες διαφημίσεις οι οποίες μπορούν να κατηγοριοποιηθούν α) σε μια διαφήμιση υπάρχοντος διαφημιστή και β) σε διαφημίσεις ενός καινούργιου διαφημιστή. Για τις διαφημίσεις αυτές χρησιμοποιείται μια σταθερή συχνότητα συμμετοχής. Για τις

υπάρχουσες διαφημίσεις, η παράμετρος γL τίθεται να είναι η μέση τιμή της συχνότητας συμμετοχής των διαφημίσεων που υπάρχουν: $\gamma_L = \frac{\sum_{L_A \epsilon L_A} \gamma_{LA}}{|L_A|}$, όπου το L_A είναι το σύνολο όλων των διαφημίσεων του διαφημιστή A, το οποίο περιέχει και την διαφήμιση L. Έτσι, αντί να υπολογίσουμε το γ_L με την παραπάνω εξίσωση, μπορούμε να χρησιμοποιήσουμε μια σταθερή τιμή για το γ_L.

4.3.6 Ανάπτυξη μεγάλης κλίμακας

Στο σημείο αυτό, γίνεται αναφορά σε κάποια σημεία που προκύπτουν με την εφαρμογή της μεθόδου μας σε πραγματικά και μεγάλης κλίμακας προβλήματα του συστήματος sponsored search. Η μέθοδος που ακολουθείται γίνεται σε δύο φάσεις: 1) την δοκιμή 2) την πρόβλεψη. Για το κομμάτι της δοκιμής, γίνεται χρήση υπό συνθήκη κατανομής πιθανότητας για κάθε ακμή του δικτύου Bayes. Στη συνέχεια, για την πρόβλεψη, δοσμένης μιας διαφήμισης L, προσομοιώνουμε το δίκτυο Bayes για να παράγουμε δημοπρασίες, στις οποίες υπολογίζουμε τον αριθμό των εμφανίσεων. Γίνεται χρήση του πλαισίου MapReduce για τον υπολογισμό του CPT από τα αρχεία δεδομένων (log files). Για παράδειγμα, έστω ότι θέλουμε να υπολογίσουμε το εξής CPT: P(vi|Fi), όπου το Fi = {vj|vi είναι παιδί του vj}. Στην αρχή, βρίσκουμε όλες τις μοναδικές τιμές που το j ∈ Fi μπορεί να πάρει. Στη συνέχεια, γίνεται μείωση (Reduce) σε κάθε συνδυασμό μοναδικών τιμών των κόμβων στο Fi και βρίσκουμε την πιθανοτική κατανομή του vi. Μετά, τα CPTs που έχουμε υπολογίσει χρησιμοποιούνται για την κατασκευή της αθροιστικής συνάρτησης κατανομής για κάθε μεταβλητή κόμβου. Έτσι, απλοποιείται και επιταχύνεται η διαδικασία της δειγματοληψίας.

Σε δεύτερη φάση, ασχολούμαστε με τη φάση πρόβλεψης της μεθόδου. Το βήμα αυτό γίνεται online και έτσι απαιτείται απάντηση σε πραγματικό χρόνο. Συνεπώς, δεν μπορούμε να προσομοιώσουμε το δίκτυο Bayes για να παράγουμε τον αριθμό των εμφανίσεων. Αντί αυτού, υπολογίζουμε ξανά τον αριθμό των εμφανίσεων σε όλες τις πιθανές τιμές προσφοράς. Κατά αυτόν τον τρόπο, όταν ένας διαφημιστής ζητά κάποια πρόβλεψη, εκτελούμε μια απλή αναζήτηση στον προβλεπόμενο αριθμό εμφανίσεων στην παρεχόμενη τιμή προσφοράς. Αν μια διαφήμιση δεν βρίσκεται στα αρχεία, αλλά ο διαφημιστής είναι παρόν, τότε μπορούμε να υπολογίσουμε τον αριθμό των εμφανίσεων για τον διαφημιστή σε όλες τις πιθανές τιμές προσφοράς και για όλες τις πιθανές λέξεις κλειδιά. Αν και ο διαφημιστής δεν είναι παρόν στα αρχεία, τότε απλά αποθηκεύουμε τον αριθμό των εμφανίσεων για κάθε λέξη κλειδί.

4.4 Πρόβλεψη του CTR με χρήση υβριδικών μοντέλων (Click-through prediction for sponsored search advertising with hybrid models)

Θα παρουσιαστεί μια προσέγγιση για τον υπολογισμό του CTR (click through rate) των διαφημίσεων. Είναι σημαντικό για την πρόβλεψη του CTR να αποφασιστούν οι τιμές των κλικς και η σειρά των εμφανίσεων. Πρώτα, υλοποιούμε τρείς μεθόδους που υπάρχουν ήδη συμπεριλαμβανομένων των BPR (Online Bayesian Probit Regression), SVM (Support Vector Machine) και το LFM (Latent Factor Model). Στη συνέχεια, προκειμένου να εκμεταλλευτούμε πλήρως το σύνολο δοκιμών, μέθοδοι της κατηγορίας MLE (Maximum

Likelihood Estimation) προτείνονται, ώστε να μοντελοποιήσουν τα στιγμιότυπα που εμφανίζονται συχνά στο σύνολο δοκιμών. Κάθε ένα από τα μοντέλα χωριστά βελτιστοποιείται, επιλέγοντας τα πιο περιγραφικά χαρακτηριστικά. Τέλος, προτείνεται μια μέθοδος που βασίζεται στην βαθμολόγηση, η οποία βελτιώνει τα αποτελέσματα του μοντέλου και τελικά η τελική έκδοση του μοντέλου βασίζεται στα μοντέλα BPR, SVM και MLE.

4.4.1 Χαρακτηριστικά

Προτείνουμε να χρησιμοποιήσουμε δύο χαρακτηριστικά για τα μοντέλα. Το πρώτο, βγαίνει απευθείας από τα δεδομένα δοκιμών, τα οποία τα ονομάζουμε αυθεντικά χαρακτηριστικά, ενώ το δεύτερο είναι τα συνθετικά χαρακτηριστικά. Στη συνέχεια, θα αναφέρουμε κάποια επιπλέον πληροφορία για τα δύο αυτά χαρακτηριστικά. Όσον αφορά τα αυθεντικά χαρακτηριστικά, το σύνολο αυτό περιέχει διακριτά και συνεχή χαρακτηριστικά. Τα διακριτά χαρακτηριστικά είναι: το μοναδικό ID κάθε διαφήμισης, ο διαφημιστής, η λέξη κλειδί, το αίτημα του χρήστη, ο τίτλος, η περιγραφή, ο χρήστης, η ηλικία του, η θέση της διαφήμισης, και το προβαλλόμενο url. Τα συνεχή χαρακτηριστικά είναι: η τιμή του click through rate (CTR) καθενός από τα διακριτά χαρακτηριστικά. Για τα σύνθετα χαρακτηριστικά κάνουμε το εξής, πρώτα συνδέουμε δύο οποιαδήποτε αυθεντικά διακριτά χαρακτηριστικά και τα χρησιμοποιούμε σαν σύνθετα χαρακτηριστικά. Επίσης, προσθέτουμε πληροφορία που αφορά τη θέση των αυθεντικών διακριτών χαρακτηριστικών. Τέλος, υιοθετούνται χαρακτηριστικά για την ανάλυση των τίτλων, των αιτημάτων του χρήστη, αλλά και περιγραφών.

4.4.2 Ατομικές μέθοδοι

Στο σημείο αυτό, παρουσιάζουμε τα μεμονωμένα μοντέλα που χρησιμοποιούμε.

4.4.2.1 Online Bayesian Probit regression

Εισήχθη από τους ερευνητές της Microsoft. Χρησιμοποιούνταν για να προβλέπει το CTR για το σύστημα διαφήμισης sponsored search στη μηχανή αναζήτησης Bing και γενικά δούλεψε πολύ καλά. Το μοντέλο διατηρεί τις γκαουσιανές πεποιθήσεις του πάνω στο βάρος του μοντέλου και επιτελεί γκαουσιανές αναβαθμίσεις που προέρχονται από την ανταλλαγή μηνυμάτων.

4.4.2.2 Η διαδικασία LDA (Latent Dirichlet Allocation)

Οι όροι στα δεδομένα δοκιμής είναι λέξεις και μια ακολουθία όρων μπορούμε να την χειριστούμε σαν ένα έγγραφο. Ως εκ τούτου, χρησιμοποιούμε τη μέθοδο LDA, για να διαχειριστούμε τα αρχεία όρων. Υποθέτουμε την ακόλουθη παραγωγική διαδικασία για μια ακολουθία όρων d: α) Επιλέγουμε θ_d ~ Dir(a), μια κατανομή Dirichlet με παράμετρο α. β) Για κάθε έναν από τους όρους N μέσα στο d: i) επιλογή ενός θέματος z_n ~ Multinomial(θd), μια κατανομή της παραμέτρου θ_d. ii) επιλογή ενός όρου t_n από το p ($t_n|z_n$, β), μια πιθανότητα εξαρτώμενη από το θέμα z_n. Για το σώμα D όλων των ακολουθιών των όρων, μπορούμε να αποκτήσουμε την πιθανότητα p(D|α,β) = $\prod_{d=1}^{M} \int p(\theta d|\alpha)$ $(\prod_{n=1} N_d \ \sum_{z_{d,n}} p(z_{d,n}|\theta_d) p(t_{d,n}|z_{d,n,\beta}))d\theta_d$.

Χρησιμοποιούμε την μέθοδο του Gibbs για να συμπεράνουμε και να αποκτήσουμε μια παράμετρο διανύσματος θd για μια ακολουθία όρων d. Χρησιμοποιούμε το διάνυσμα θd σαν τα χαρακτηριστικά της ακολουθίας των όρων d.

4.4.2.3 Χαρακτηριστικά

Εφαρμόζουμε την μέθοδο LDA για να γίνει μια προεπεξεργασία των όρων του αιτήματος του χρήστη, των όρων των λέξεων κλειδιά και των όρων του τίτλου. Επίσης, χαρακτηριστικά όπως το ID, πληροφορίες για τον χρήστη και ομοιότητες μεταξύ του αιτήματος του χρήστη και τη διαφήμιση συμπεριλαμβάνονται. Χρησιμοποιούνται τριών ειδών ομοιότητες: η ομοιότητα μεταξύ αιτήματος χρήστη και αγορασμένης λέξης κλειδί (sim_{q_k}), η ομοιότητα μεταξύ αιτήματος χρήστη και τίτλου διαφήμισης (sim_{q_t}) και η ομοιότητα μεταξύ αιτήματος χρήστη και περιγραφής της διαφήμισης (sim_{q_d}). Τα αναπαριστούμε με τρία δυδιάστατα διανύσματα: (sim_{q_k} , 1 - sim_{q_k}), (sim_{q_t} , 1 - sim_{q_t}), (sim_{q_l} , 1 – sim_{q_l}). Το τελικό διάνυσμα χαρακτηριστικών είναι το x και αποτελείται από πολλές ομάδες διανυσμάτων: $x = (x1^T, x2^T, \ldots, xN^T)^T$ όπου $xi = \begin{pmatrix} x_{i,1} \\ x_{i,2} \\ x_{i,Mi} \end{pmatrix}$, $\sum_{j=1}^{Mi} x_{i,j} = 1$.

4.4.3 Μοντέλο πιθανότητας και γράφημα παραγόντων

Το μοντέλο Bayes Probit Regression βασίζεται σε ένα γενικότερο γραμμικό μοντέλο με μια συνάρτησης σύνδεσης : $p(y|x,w) = \Phi(\frac{y\, w^T x}{\beta})$, όπου το $\Phi(t) = \int_{-\infty}^{t} N(s: 0,1)ds$ είναι η τυποποιημένη σωρευτική γκαουσιανή πυκνότητα, η οποία εξυπηρετεί σαν μια αντίστροφη συνάρτηση που αντιστοιχίζει την έξοδο του γραμμικού μοντέλου από το (- ∞, ∞) στο (0,1). Το x δείχνει το διάνυσμα χαρακτηριστικών και το w είναι ένα διάνυσμα με τα βάρη. Το διάνυσμα w κατηγοριοποιείται σύμφωνα με το διάνυσμα χαρακτηριστικών x: $w = (w1^T, w2^T, \ldots, wN^T)^T$. Το y ∈ {-1,+1}, όπου το -1 αναπαριστά ένα μη κλικάρισμα, και το +1 αναπαριστά ένα κλικ. Η παράμετρος β δείχνει την κλίση της αντίστροφης σύνδεσης της συνάρτησης. Προκειμένου να φθάσουμε σε έναν online αλγόριθμο Bayes, βασιζόμενοι στο παραπάνω μοντέλο, εισάγουμε δύο μεταβλητές: $p(w) = \prod_{i=1}^{N} \prod_{1}^{Mi} N\left(w_{i,j}, \mu_{i,j}, {\sigma_{i,j}}^2\right)$,

$S = w^T x$, $p(s) = N(s|x^T\mu, (x^2)^T\sigma^2)$

$P(t|s) = N(t, s, \beta 2)$ και $p(y|t) = \Phi(yt)$.

4.4.4 Μοντέλο παραγόντων (Latent Factor Model)

Στο μοντέλο αυτό, οι χρήστες και τα αντικείμενα προβάλλονται σε έναν παράγοντα διάστασης K και η πρόβλεψη υπολογίζεται από το εσωτερικό γινόμενο του διανύσματος χαρακτηριστικού p_u και του διανύσματος χαρακτηριστικού q_i, $\widetilde{rui} = p_u q_i$, όπου το p_u και το q_i μπορούμε να τα βρούμε ελαχιστοποιώντας την ακόλουθη συνάρτηση: $\min\sum(r_{ui} - \widetilde{r_{ui}}) + \lambda(||p_u||^2 + ||q_i||^2)$. Η πρόβλεψη του CTR μπορεί να μετατραπεί σε ένα πρόβλημα αξιολόγησης. Αν ο χρήστης κλικάρει πάνω σε μια διαφήμιση, μπορούμε να χαρτογραφήσουμε τον χρήστη και την διαφήμιση σε ένα χώρο παραγόντων. Έχουμε περισσότερη πληροφορία για τον χρήστη, τον τίτλο της διαφήμισης, την περιγραφή, τον

διαφημιστή και το αίτημα του χρήστη. Για τους όρους, επιλέγουμε τους N μεγαλύτερους για κάθε διαφήμιση διαμέσου του TFIDF. Για τους καινούργιους χρήστες, αντιστοιχίζουμε τις δημογραφικές πληροφορίες του χρήστη, όπως την ηλικία και το φύλο σε έναν παράγοντα χώρου για να περιγράψουμε αυτό το είδος των χρηστών. Το ακριβές μοντέλο είναι αυτό που ακολουθεί:

$\widetilde{rui}$ = μ + $\sum_{fi \in Fi} bfi$ + $\sum_{fu \in Fu} bfu$ +($\sum_{fi \in Fi} qfi$)^T ($\sum_{fu \in Fu} pfu$).

Το Fi δείχνει το σύνολο όλων των χαρακτηριστικών, ενώ το Fu είναι το σύνολο όλων των χαρακτηριστικών των χρηστών. Όμοια, μπορούμε να μάθουμε τις παραμέτρους, χρησιμοποιώντας μια στοχαστική μέθοδο βελτιστοποίησης για να ελαχιστοποιήσουμε την τετραγωνική συνάρτηση λάθους. Η πολυπλοκότητα του προβλήματος είναι γραμμική με τον αριθμό των εμφανίσεων. I = min $\sum(r_{ui} - \widetilde{r_{ui}})^2$ + λ ($\sum_{fi \in Fi} ||bfi||^2$ + $\sum_{fu \in Fu} ||bfu||^2$ + $\sum_{fi \in Fi} ||qfi||^2$ + $\sum_{fu \in Fu} ||pfu||^2$).

4.4.5 Feature-based Maximum Likelihood Estimation - MLE

Η μέθοδος αυτή χρησιμοποιείται προκειμένου να μοντελοποιηθεί η επιρροή διαφορετικών χαρακτηριστικών. Για παράδειγμα, υπολογίζουμε το click through rate (CTR) για κάθε χαρακτηριστικό. Χρησιμοποιούμε επίσης, τα αυθεντικά διακριτά χαρακτηριστικά καθώς και τα χαρακτηριστικά θέσης. Το CTR των διαφημίσεων μειώνεται σημαντικά και έτσι έχουμε το εξής: Clicks Over Expected Clicks - COEC = $\frac{\sum_{r=1}^{R} C_r}{\sum_{r=1}^{R} i_r \, x \, CTR_r}$. Ο αριθμητής είναι ο συνολικός αριθμός των κλικ που λαμβάνεται από μια μεταβλητή χαρακτηριστικού και ο παρονομαστής μπορεί να ερμηνευθεί σαν τα αναμενόμενα κλικ που μια μέση διαφήμιση μπορεί να λάβει, αφού παρουσιαστεί i_r φορές στην κατάταξη r. Το CTR_r είναι το μέσο CTR για κάθε θέση στη σελίδα αποτελεσμάτων.

4.4.6 Σταθμισμένο σύνολο ξεχωριστών χαρακτηριστικών

Από τη στιγμή που έχουμε όλα τα χαρακτηριστικά, σκοπός είναι να αποκτήσουμε το CTR για κάθε χαρακτηριστικό. Χρησιμοποιούμε το f_i για να δείξουμε το χαρακτηριστικό i και το f_{CTRi} για να δείξουμε το CTR του f_i. Έτσι, υπολογίζουμε το ποσοστό των θετικών στιγμιότυπων σε όλα τα στιγμιότυπα που περιέχουν το f_i. Με χρήση των χαρακτηριστικών, ξεχωριστά μπορούμε να πάρουμε την τιμή AUC, για κάθε χαρακτηριστικό του συνόλου. Η τιμή αυτή w_i χρησιμοποιείται σαν το βάρος του f_i. Προκειμένου να ανεβάσουμε τις πληροφορίες που αφορούν τη θέση, υπολογίζουμε ξεχωριστά το f_{CTR} σύμφωνα με τις θέσεις τους:

$f_{CTRi,j}$ = f_{CTRi} x $\frac{CTRj}{\sum_{r=1}^{|R|} CTRr}$, όπου το $f_{CTRi,j}$ είναι το CTR του χαρακτηριστικού i στη θέση j και το f_{CTRi} είναι το CTR του χαρακτηριστικού i. Το CTRj εκφράζει το CTR των διαφημίσεων που έχουν τη θέση j και το R είναι το σύνολο των θέσεων. Όταν προβλέπουμε το CTR ενός στιγμιότυπου, πρώτα παράγουμε το σύνολο χαρακτηριστικών αυτού. Υποθέτοντας ότι το F είναι μια συλλογή από στιγμιότυπα χαρακτηριστικών, το CTR υπολογίζεται, όπως στη συνέχεια: CTR = $\sum_{i=1}^{|F|} (f_{CTRi} \, x \, w_i)$.

4.5 Εξόρυξη πληροφορίας για τη συμπεριφορά του χρήστη (Mining advertiser specific user behavior using Adfactors)

Θα περιγραφεί ένα εναλλακτικό μοντέλο εξόρυξης πληροφορίας για την ανάλυση σε επίπεδο χρήστη των δεδομένων των διαφημίσεων. Ιστορίες χρηστών συμπιέζονται σε ένα γράφημα, που το ονομάζουμε adgraph. Αυτό αναπαριστά τοπικές σχέσεις ανάμεσα σε διαφημιστικά γεγονότα. Για το βήμα της αναφοράς, εισάγουμε κάποιους κανόνες βαθμολόγησης, οι οποίοι ονομάζονται adfactors (AF), και μπορούν να συλλάβουν τον συνολικό ρόλο των διαφημίσεων και των μονοπατιών στο γράφημα διαφημίσεων (adgraph) και πιο συγκεκριμένα, η δομική συσχέτιση στην εμφάνιση μιας διαφήμισης και στη μετατροπή του χρήστη. Παρουσιάζουμε, τοπικούς αλγόριθμους για τον υπολογισμό των παραγόντων αυτών (adfactors). Όλοι οι αλγόριθμοι υλοποιήθηκαν με χρήση του μοντέλου προγραμματισμού MapReduce και του πλαισίου Pregel.

4.5.1 Ad factors

Στο σημείο αυτό, ορίζουμε και αξιολογούμε έναν αριθμό από adfactors, που δεσμεύουν τη δομική σχέση στην μετατροπή του μονοπατιού πληροφορίας. Όλοι οι adfactors έχουν υπολογιστεί απευθείας στις δομές του γραφήματος. Κάποιες από τις δομές αυτές μπορεί να είναι για παράδειγμα ένα γράφημα Markov. Σε αυτό, κάθε κόμβος του γραφήματος είναι ένα γεγονός, δηλαδή μια εμφάνιση ή ένα κλικ και τα βάρη των ακμών είναι απλά πιθανότητες από την παρατήρηση δύο διαδοχικών γεγονότων σε ένα μονοπάτι. Παρουσιάζουμε επίσης, διαφορετικές εκδοχές adfactors και αξιολογούμε τις ιδιότητες τους πάνω σε πραγματικά δεδομένα. Ο παράγοντας αυτός (adfactor), υπολογίζεται για κάθε κόμβο του γραφήματος και άτυπα μπορεί να θεωρηθεί σαν ένα μέτρο συσχέτισης του κόμβου αυτού σε σχέση με τον κόμβο στόχο. Ο πιο απλός adfactor μπορεί να είναι μια απλή βεβαρημένη ακμή ανάμεσα σε ένα ζευγάρι κόμβων του γραφήματος.

Στη συνέχεια, παρουσιάζονται τρία σημαντικά σημεία των παραγόντων αυτών. Το πρώτο είναι το εξής: ενώ περιμένουμε οι παράγοντες να έχουν χρήσιμες εφαρμογές και να χρησιμεύουν στην κατάταξη και στην υποβολή των προθέσεων μας, δεν μπορούμε να επισυνάψουμε καμία έννοια αιτίας στις τιμές. Σε δεύτερο στάδιο, δεν υπάρχει κάποιο μέτρο ποιότητας του παράγοντα αυτού και δεν μπορούμε να πούμε με ακρίβεια ότι ένας παράγοντας είναι καλύτερος από κάποιον άλλο. Τρίτον, προκειμένου οι παράγοντες να είναι πρακτικά χρήσιμοι, θα πρέπει να είναι υπολογίσιμοι σε μεγάλα σύνολα πληροφορίας.

4.5.2 Ο παράγοντας LastAd

Ο παράγοντας αυτός παρουσιάζει τον βαθμό αντίθεσης της αντίστοιχης διαφήμισης. Μετρά το βάρος της απευθείας ακμής από τον αντίστοιχο κόμβο προς τον κόμβο μετατροπής. Πρακτικά, για οποιοδήποτε γεγονός e, είτε απλή εμφάνιση διαφήμισης, είτε κλικ πάνω σε κάποια διαφήμιση. Έτσι, ο παράγοντας LastAd είναι:

$$\text{Lastad (e)} = \frac{αριθμός\ εμφανίσεων\ του\ \{e,conversion\}}{αριθμός\ εμφανίσεων\ του\ e}$$

4.5.3 Η συνεισφορά του PageRank (PageRank contribution PPR, CPR)

Ορίζεται σαν μια συνεισφορά του κόμβου v στον κόμβο μετατροπής c (ppr(v,c)). Στη συνέχεια, δίνεται ένας πιο επίσημος ορισμός. Υποθέτουμε μια πιθανότητα επανεκκίνησης, έστω α > 0 και ένα γράφημα που δίνεται από τον πίνακα M μεγέθους n x n. Έστω I ο ταυτοτικός πίνακας και e_v είναι ένα μοναδιαίο διάνυσμα γραμμής, του οποίου η v καταχώρηση ισούται με ένα.

Ο PPR (Personalized PageRank) πίνακας ορίζεται σαν ένας n x n πίνακας. Ακολουθεί η εξίσωση: $ppr_\alpha = \alpha I + (1-\alpha)ppr_\alpha$ M. Φτιάχνουμε έναν κόμβο v. Η v γραμμή του πίνακα ικανοποιεί το εξής : $ppr_\alpha(v,) = \alpha e_v + (1-\alpha)ppr_\alpha(v,)$ M. Αυτό είναι γνωστό σαν PPR διάνυσμα ενός κόμβου v, με πιθανότητα επανεκκίνησης α. Η διπλή κατασκευή του διανύσματος αποκτάται παίρνοντας μια μοναδική στήλη του πίνακα. Η στήλη που αντιστοιχεί στον κόμβο μετατροπής c, αποτελείται από όλες τις PPR συνεισφορές για διαφορετικούς κόμβους v και για το σταθερό κόμβο μετατροπής c. Στη συνέχεια, λύνει την ακόλουθη εξίσωση: $ppr_\alpha(,c) = \alpha e_c + (1-\alpha)ppr_\alpha(,c)\ M^T$. Το διάνυσμα αυτό δεσμεύει σημαντικές δομικές ιδιότητες του γραφήματος, όπως είναι ο αριθμός των διαφορετικών μονοπατιών από κάθε κόμβο v στον κόμβο μετατροπής c, όπως επίσης και το μήκος των μονοπατιών αυτών. Μια ωραία ιδιότητα του cpr είναι ότι το άθροισμα όλων των cpr του κόμβου c ισούται με το PageRank του κόμβου c, ώστε να θεωρείται σαν ένας τρόπος να διαιρούμε τη βαθμολογία του PageRank σε συνεισφορές και άλλων κόμβων.

4.5.4 Αποτελεσματικοί αλγόριθμοι (Efficient Algorithms)

Η αποτελεσματική ικανότητα υπολογισμού είναι σημαντική για την επιτυχή εφαρμογή οπουδήποτε αλγόριθμου data mining σε πραγματικές διαφημίσεις. Το κλειδί είναι η παρατήρηση ότι τα διανύσματα της συνεισφοράς του PageRank μπορούν να υπολογιστούν με χρήση ενός τοπικού αλγόριθμου, που εξετάζει μόνο ένα μικρό κομμάτι του γραφήματος εισόδου δίπλα σε ένα συγκεκριμένο διάνυσμα. Ο υπολογισμός του cpr μπορεί να επιτευχθεί από έναν απλό τοπικό αλγόριθμο, στον οποίο κάθε κόμβος αποτελείται από δύο μεταβλητές. Το τρέχον υπολογισμένο cpr, και η τιμή που έχει αποκτηθεί από άλλους κόμβους. Οι κόμβοι τρέχουν μια απλή διαδικασία μέχρι η τιμή του κάθε κόμβου να μην ξεπερνάει μια τιμή ε.

4.5.4.1 Ο αλγόριθμος

Algorithm 1 Local algorithm to calculate $\text{cpr} \equiv \text{ppr}(\cdot, c)$ adfactor of the conversion node c.

Initialization:
$\text{cpr}(u) \Leftarrow 0$, $\text{resid}(u) \Leftarrow 1$ if conversion node otherwise 0
Main Loop:
while $\exists u$ such that $|\text{resid}(u)| \geq \varepsilon$ **do**
Pushback(u, α)
end while
Pushback (u, α)**:**
$\text{cpr}(u) \Leftarrow \text{cpr}(u) + \alpha\,\text{resid}(u)$ {accumulate α fraction of the current residual}
for every incoming edge $w \longrightarrow u$ **do**
$\text{resid}(w) \Leftarrow \text{resid}(w) + (1-\alpha)\frac{\text{resid}(u)}{d_{out}(w)}$ {distribute $1-\alpha$ fraction of the current residual to neighbors}
end for
$\text{resid}(u) \Leftarrow 0$

4.5.4.2 Θεώρημα 1

Ο παραπάνω αλγόριθμος υπολογίζει μια ε-προσέγγιση της συνεισφοράς PageRank για τον κόμβο μετατροπής. Για παράδειγμα, το cpr = ppr(,c) , με χρήση μόνο (1/αε + 1) πράξεις. Επίσης, κάποιος μπορεί να αναγνωρίσει τους μεγαλύτερους k κόμβους με τη μεγαλύτερη PageRank συνεισφορά, χρησιμοποιώντας μόνο O($\frac{k}{\alpha}$) πράξεις.

4.5.4.3 Θεώρημα 2

Υπάρχει μια τοπική ε-προσέγγιση στον αλγόριθμο για τον υπολογισμό του MI_α και RE_α με χρήση O(1/αε) πράξεις. Στη συνέχεια, μια ε-προσέγγιση για το Pass, MI, και το RE μπορεί να υπολογιστεί χρησιμοποιώντας O(1/α'ε), που α' = $\min_i w_{i,null}$.

4.5.5 Εργαλεία του data mining

Στο τμήμα αυτό συζητάμε διάφορα εργαλεία data mining που μπορούν να υλοποιηθούν με χρήση παραγόντων διαφήμισης (ad factors) και ορίζονται από μοντέλα διαφημίσεων που έχουν περιγραφεί παραπάνω. Πρώτα, μπορούν να αναγνωριστούν οι k διαφημίσεις με τον μεγαλύτερο παράγοντα. Για παράδειγμα, ο παράγοντας PPR έχει υψηλή, αλλά όχι τέλεια σχέση με το χαρακτηριστικό της μάρκας σε μια εμφάνιση διαφήμισης. Έτσι, η αναγνώριση των k διαφημίσεων με τον παράγοντα PPR μπορούν να ερμηνευθούν σαν την αναγνώριση των καλύτερων διαφημίσεων με την πιο καλή επιρροή. Κάποια από τα αιτήματα των χρηστών θα περιλαμβάνουν και κάποιο όνομα διαφημιστή ή εταιρίας. Βέβαια, κάποιες μπορεί και να μην το έχουν και έτσι μπορούν να θεωρηθούν σαν μάρκα «σκιά». Ένα δεύτερο χαρακτηριστικό είναι η αναγνώριση διαφημίσεων με σημαντικά μακροπρόθεσμα αποτελέσματα, τα οποία δεν συμπεριλαμβάνονται στο μοντέλο Last Ad. Οι διαφημιστές γενικά στηρίζονται στα αποτελέσματα του μοντέλου Last ad για να εκτιμήσουν την επιρροή των διαφημίσεων. Ένα εναλλακτικό μοντέλο είναι να κοιτάζουμε για διαφημίσεις, οι οποίες βαθμολογούνται ψηλότερα με παράγοντες που λαμβάνουν υπόψη τους μακροπρόθεσμες σχέσεις, όπως το PPR. Ένας τρόπος κατάταξης είναι η διαφορά ανάμεσα στον παράγοντα PPR και στον παράγοντα Last Ad. Ένα άλλο κριτήριο είναι να κοιτάζουμε για εμφανίσεις που έχουν τον παράγοντα PPR να μεγαλώνει με την απόσταση της μετατροπής ή να είναι ο παράγοντας PPR στην κατάσταση ψαξίματος (search state), υψηλότερα από την κατάσταση του ενδιαφέροντος (interested state).

4.5.6 Επεξήγηση του παράγοντα ad factor

Ο παράγοντας που εκχωρείται σε κάθε κόμβο του γραφήματος, θα πρέπει να επεξηγείται εύκολα από μια τοπική δομή γύρω από αυτόν τον κόμβο. Για παράδειγμα, για τον παράγοντα PPR ενός κόμβου u, κάποιος θα μπορεί πάντα να σκεφτεί τους μεγαλύτερους m γειτονικούς κόμβους διαμέσου των οποίων ο κόμβος u συνεισφέρει τη μεγαλύτερη ποσότητα στο PageRank στον κόμβο μετατροπής. Τυπικά, ας θωρήσουμε ότι το π_u είναι η ανταλλαγή όλων των γειτόνων του κόμβου u που τους κατατάσσει κατά φθίνουσα σειρά, δηλαδή $w_{uv}ppr_\alpha(u,c)$. Ορίζουμε το $\pi_u(m)$ σαν το σύνολο των πρώτων m κόμβων της ανταλλαγής. Αυτοί οι κόμβοι μπορούν να θεωρηθούν σαν τις πιο πιθανές επόμενες πράξεις του χρήστη.

Βιβλιογραφία κεφαλαίου

[1] Zeyuan Allen Zhu, Weizhu Chen, Tom Minka, Chenguang Zhu, Zheng Chen. “A Novel Click Model and Its Applications to Online Advertising”. In *Proceedings of the third ACM international conference on Web search and data mining, (WSDM 2010).* New York City, USA. February 4-6, 2010.

[2] Eric Sodomka, Sebastien Lahaie, Dustin Hillard. “A Hierarchical Model for Value Estimation in Sponsored Search”. In *EC 2011 Seventh Ad Auctions Workshop*. June 2011.

[3] Abhirup Nath, Shibnath Mukherjee, Prateek Jain, Navin Goyal, Srivatsan Laxman. “Ad Impression Forecasting for Sponsored Search”. In *Proceedings of the 22nd international conference on World Wide Web Pages 943-952(WWW 2013).* Rio de Janeiro, Brazil. May 13-17, 2013.

[4] Xingxing Wang, Shijie Lin, Dongying Kong, Liheng Xu, Qiang Yan, Siwei Lai, Liang Wu, Alvin Chin, Guibo Zhu, Heng Gao, Yang Wu, Danny Bickson, Yuanfeng Du, Neng Gong, Chengchun Shu, Shuang Wang, Kang Liu, Shuren Li, Jun Zhao, Fei Tan, Yuanchun Zhou. “Click-Through Prediction for Sponsored Search Advertising with Hybrid Models”. *ACM KDD CUP Workshop*, 2012.

[5] Nikolay Archak, Vahab S. Mirrokni, S. Muthukrishnan. “Mining Advertiser-specific User Behavior Using Adfactors”. In *Proceedings of the 19th international conference on the World Wide Web, Pages 41-40(WWW 2010)*. Raleigh, North Carolina, USA. April 26-30, 2010.

[6] Panagiotis Papadimitriou. “Algorithms and Strategies for Web Advertising”(Chapter 2). Stanford University, Dept. of Electrical Engineering. December 2011

Κεφάλαιο 5: Display Advertising and Delivery

5.1 Εννοιολογική προσέγγιση της contextual διαφήμισης (A semantic approach to contextual advertising)

Η έννοια του contextual advertising ή contextual match (CM) αναφέρεται στην τοποθέτηση εμπορικού κειμένου διαφημίσεων μέσα στο περιεχόμενο μιας ιστοσελίδας. Σε αντίθεση με το σύστημα του Sponsored Search (SS), που μελετήσαμε σε παραπάνω κεφάλαιο, όπου έχουμε την τοποθέτηση διαφημίσεων στη σελίδα αποτελεσμάτων, με τις διαφημίσεις να προέρχονται από κάποιο αίτημα του χρήστη. Στο CM υπάρχει συνήθως μια ενδιάμεση οντότητα, δηλαδή ένα δίκτυο διαφημίσεων, το οποίο είναι υπεύθυνο να βελτιστοποιεί την επιλογή των διαφημίσεων και ταυτόχρονα να καλύπτει και την εμπειρία και τις ανάγκες του χρήστη.

Οι περισσότερες διαφημίσεις κειμένου χαρακτηρίζονται από λέξεις προσφορών (bid phrases), όπως στο σύστημα sponsored search (SS) και αναπαριστούν εκείνα τα αιτήματα του χρήστη, όπου οι διαφημιστές θα ήθελαν οι διαφημίσεις τους να εμφανίζονται. Σε πρώτη φάση, οι πρώτες τεχνολογίες που χρησιμοποιήθηκαν στο contextual matching (CM), ήταν να εξαχθούν μία ή περισσότερες φράσεις από το περιεχόμενο της σελίδας, και να παρουσιαστούν διαφημίσεις που να αντιστοιχούν σε αναζήτηση πάνω στις φράσεις αυτές, με μια απλή συντακτική προσέγγιση. Παρόλα αυτά, κάτι τέτοιο οδηγεί σε αρκετές μη σχετικές διαφημίσεις. Προκειμένου να αντιμετωπιστεί αυτή η κατάσταση, προτείνεται ένα σύστημα για διαφημίσεις τύπου contextual, βασισμένο στο συνδυασμό εννοιολογικών και συντακτικών χαρακτηριστικών.

5.1.1 Επισκόπηση της contextual διαφήμισης (Overview of contextual advertising)

Η διαφήμιση αυτή (contextual advertising) είναι μια αλληλεπίδραση τεσσάρων παραγόντων:

- *Ο εκδότης* : είναι ο ιδιοκτήτης των ιστοσελίδων, στις οποίες θα παρουσιαστεί η διαφήμιση. Ουσιαστικά, σκοπός του είναι να μεγιστοποιήσει τα έσοδα από τη διαφήμιση, ενώ ταυτόχρονα θα προσφέρει και μια καλή εμπειρία στο χρήστη.
- *Ο διαφημιστής* : παρέχει τον ανεφοδιασμό των διαφημίσεων. Συχνά, η δραστηριότητα των διαφημιστών οργανώνεται γύρω από καμπάνιες, οι οποίες ορίζονται από ένα σύνολο διαφημίσεων με ένα στόχο. Σκοπός των διαφημιστών είναι να προωθήσουν προϊόντα και υπηρεσίες.
- *Ένα δίκτυο διαφημίσεων* : είναι ο διαμεσολαβητής ανάμεσα στον διαφημιστή και στον εκδότη και επιλέγει διαφημίσεις, οι οποίες έχουν τοποθετηθεί στις σελίδες. Το δίκτυο αυτό μοιράζεται τα έσοδα των διαφημίσεων με τον εκδότη.
- *Οι χρήστες :* επισκέπτονται τις ιστοσελίδες και αλληλεπιδρούν με τις διαφημίσεις.

Το contextual advertising, συχνά μπορεί να ερμηνευθεί και σαν άμεσο marketing, όπου γενικά σκοπός είναι η «άμεση απάντηση», όπου η επιρροή μια καμπάνιας μετράται από την

αντίδραση του χρήστη. Ένα από τα πλεονεκτήματα της online διαφήμισης είναι ότι ο χρήστης μπορεί να ακολουθήσει το link που βρίσκεται μέσα σε μια διαφήμιση και έτσι να επισκεφθεί την ιστοσελίδα του διαφημιστή. Αυτό είναι και το οικονομικό μοντέλο, όπου ο διαφημιστής πληρώνει ένα συγκεκριμένο χρηματικό ποσό για κάθε κλικ πάνω στη διαφήμιση.

Σε ένα δίκτυο διαφημίσεων έχουμε αντιστοιχία ενδιαφερόντων του εκδότη, του διαφημιστή και του δικτύου. Γενικά, τα κλικ φέρνουν έσοδα και στον εκδότη και στο δίκτυο διαφημίσεων, αλλά και στον διαφημιστή, φέρνοντας κίνηση στην επισκεψημότητα της ιστοσελίδας. Τα έσοδα ενός δικτύου, δοσμένης μιας σελίδας p, μπορούν να εκτιμηθούν σαν R = $\sum_{i=1\ldots k} P(click|p,a_i)price(a_i,i)$, όπου το k είναι ο αριθμός των διαφημίσεων που εμφανίζονται στη σελίδα p και το price(a_i,i), είναι η τιμή για κάθε κλικ της τρέχουσας διαφήμισης α_i στη θέση i. Η τιμή του μοντέλου εξαρτάται από το σύνολο των διαφημίσεων που παρουσιάζονται στη σελίδα. Γενικά, επικεντρωνόμαστε στο να βρίσκουμε διαφημίσεις που να μεγιστοποιούν το πρώτο στοιχείο του προϊόντος. Δηλαδή, ψάχνουμε για argmax(Pclick|p,a_i).

5.1.2 Μέθοδος κατάταξης

Η εννοιολογική φάση του ταιριάσματος βασίζεται σε διαφημίσεις και σελίδες που είναι τοπικά κοντά. Έτσι, στο σημείο αυτό θα μελετήσουμε κάποιες μεθόδους που χρησιμοποιούνταν για να κατασκευαστεί μια σελίδα, αλλά και ένα ζεύγος κατάταξης διαφημίσεων. Σε δύο από τις μεθόδους αυτές, προσπαθήσαμε να βρούμε σελίδες σαν δοκιμή, αλλά και διαφημίσεις για κάθε περιεχόμενο σελίδας, όπως παρακάτω: υλοποιήσαμε ένα σύνολο δοκιμών για τις σελίδες, τρέχοντας τις ερωτήσεις του χρήστη στην ταξινόμηση και χρησιμοποιώντας τα πιο υψηλά δέκα αποτελέσματα και μετά από κάποιου είδους φιλτράρισμα στα έγγραφα, τα οποία επισημαίνονται με την σήμανση της ερώτησης. Από την πλευρά των διαφημίσεων, παράγαμε ένα σύνολο δοκιμών για κάθε τάξη, επιλέγοντας διαφημίσεις οι οποίες έχουν μια φράση προσφοράς που να αντιστοιχεί στην τάξη αυτή. Με τα σύνολα αυτά, εξετάσαμε μια ιεραρχία SVM και μια κατάταξη με τη μέθοδο της οπισθοδρόμησης. Παρόλα αυτά, αποκτήσαμε την καλύτερη επίδοση, χρησιμοποιώντας το τρίτο έγγραφο κατάταξης, το οποίο βασίζεται στην πληροφορία που αποκτά από τα αιτήματα του χρήστη. Για κάθε κόμβο κατάταξης, ενώσαμε όλα τα αιτήματα σε ένα και μοναδικό έγγραφο. Στη συνέχεια, χρησιμοποιήσαμε κάθε τέτοιο έγγραφο σαν κεντρικό για κάποιο κοντινό γείτονα ταξινόμησης. Κάθε κεντρικό έγγραφο ορίζεται από το άθροισμα των τιμών tf-idf για κάθε όρο, κανονικοποιημένο από τον αριθμό των αιτημάτων της τάξης: $\overrightarrow{c_j} = \frac{1}{|C_j|}\sum_{\vec{q}\in C_j}\frac{\vec{q}}{||\vec{q}||}$ και όπου το c_j είναι ο κεντρικός της τάξης C_j, το q επαναλαμβάνεται ανάμεσα στα αιτήματα του χρήστη σε μια συγκεκριμένη τάξη.

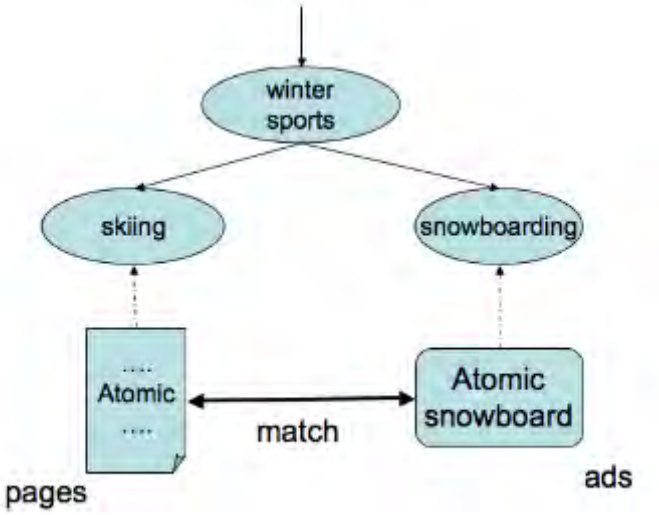
5.1.3 Σημασιολογικό – συντακτικό ταίριασμα (Semantic – syntactic matching)

Τα συστήματα του contextual advertising επεξεργάζονται το περιεχόμενο της σελίδας, εξάγουν χαρακτηριστικά, και στη συνέχεια ψάχνουν το χώρο της διαφήμισης για να βρουν τις διαφημίσεις που ταιριάζουν καλύτερα. Δοσμένης μιας σελίδας p και ένα σύνολο διαφημίσεων A = {$\alpha_1\ldots\alpha_s$}, υπολογίζουμε την πιθανότητα σχετικότητας του κλικ P(click|p,α) με μια βαθμολογία, η οποία δεσμεύει την ποιότητα ταιριάσματος της σελίδας και της

διαφήμισης. Προκειμένου να βρεθούν οι καλύτερες διαφημίσεις για μια σελίδα, ταξινομούμε τις διαφημίσεις στο σύνολο A και επιλέγουμε λίγες της πιο υψηλής κατάταξης να παρουσιάσουμε. Το πρόβλημα μπορεί να οριστεί ως η αντιστοίχιση κάθε σελίδας του συνόλου P = $\{p_1,\ldots,p_{pc}\}$ σε μία ή περισσότερες διαφημίσεις του συνόλου διαφημίσεων. Κάθε σελίδα αναπαρίσταται σαν ένα σύνολο από τμήματα σελίδων $p_i = \{p_{i,1}, p_{i,2} \ldots p_{i,m}\}$. Τα τμήματα της σελίδας αναπαριστούν διαφορετικά δομικά κομμάτια, όπως ο τίτλος, η μεταπληροφορία, το κυρίως σώμα κλπ. Με τη σειρά του, κάθε κομμάτι είναι ένα αταξινόμητο σύνολο λέξεων. Μια σελίδα αναπαρίσταται από το σύνολο των όρων σε κάθε τμήμα: pi = $\{pw_1^{s1}, pw_2^{s1} \ldots pw_m^{si}\}$, όπου το pw σημαίνει μια λέξη σελίδας και ο εκθέτης δείχνει το τμήμα κάθε όρου. Παρόμοια, αναπαριστούμε κάθε διαφήμιση σαν ένα σύνολο από τμήματα $\alpha = \{\alpha_1,\alpha_2,\ldots,\alpha_l\}$, και κάθε τμήμα με τη σειρά του είναι ένα ακανόνιστο σύνολο όρων : αi = $\{ aw_1^{s1}, aw_2^{s1} \ldots aw_l^{sj} \}$, όπου το aw είναι μια λέξη διαφήμισης.

Στη συνέχεια, κάθε σελίδα και όρος διαφήμισης σχετίζεται με ένα βάρος το οποίο βασίζεται στις τιμές tf-idf. Μια τιμή tf καθορίζεται βασιζόμενη σε ανεξάρτητα τμήματα διαφημίσεων. Υπάρχουν επίσης διαφορετικές επιλογές για την τιμή idf, βασιζόμενοι σε διαφορετική οπτική. Προκειμένου να συνδυαστεί η επιρροή του τμήματος των όρων και τα αποτελέσματα στις τιμές tf-idf, το βάρος για τους όρους μιας διαφήμισης ορίζεται ως: $tWeight(kw^{si}) = weightSection(S_i)tf\_idf(kw)$. Το tw σημαίνει term weight, δηλαδή το βάρος του κάθε όρου και το weightSection(Si) είναι το βάρος που έχει αντιστοιχιστεί σε μια σελίδα ή τμήμα διαφήμισης. Ορίζουμε τις εξής δύο αντιστοιχίσεις: $Tax(p_i) = \{pc_{i1},\ldots,pc_{iu}\}$ και $Tax(a_j) = \{ac_{j1},\ldots,ac_{jv}\}$, όπου το pc,ac είναι η τάξη της σελίδας και της τάξης αντίστοιχα. Κάθε εκχώρηση σχετίζεται με ένα βάρος. Τα βάρη αυτά κανονικοποιούνται στο άθροισμα στη μονάδα: $\sum_{c\in Tax(x_i)} cWeight(c) = 1$, όπου το x_i είναι είτε μια σελίδα, είτε μια διαφήμιση, και το cWeights(c) είναι το βάρος της τάξης. Ο αριθμός των κλάσεων μπορεί να κυμαίνεται ανάμεσα σε διαφορετικές σελίδες και διαφημίσεις. Συχνά, ο αριθμός αυτός είναι 1-4. Έπειτα, ορίζουμε τη βαθμολογία σχετικότητας της διαφήμισης α_i και της σελίδας p_i ως έναν κυρτό συνδυασμό της λέξης κλειδί και της βαθμολογίας κατάταξης: $Score(p_i,\alpha_i) = \alpha\ TaxScore(Tax(p_i), Tax(a_i)) + (1-\alpha)KeywordScore(p_i,\alpha_i)$. Η παράμετρος α ορίζει το σχετικό βάρος της βαθμολογίας κατάταξης και της βαθμολογίας της λέξης κλειδί. Προκειμένου να υπολογιστεί η βαθμολογία της λέξης κλειδί (keyword score), το μοντέλο όπου οι σελίδες και οι διαφημίσεις αναπαριστώνται από ένα n-διάστατο διάνυσμα - κάθε διάσταση για κάθε ξεχωριστό όρο. Το μέγεθος κάθε διάστασης ορίζεται από μια φόρμα της μορφής tWeight(). Η βαθμολογία της λέξης κλειδί ορίζεται σαν το συνημίτονο της γωνίας ανάμεσα στα διανύσματα της σελίδας και της διαφήμισης: $KeywordScore(p_i,\alpha_i) = \frac{\sum_{i\in|K|} tWeight(pw_i)tWeight(aw_i)}{\sqrt{\sum_{i\in|K|}(tWeight(pw_i))^2)}\sqrt{\sum_{i\in|K|}(tWeight(aw_i))^2}}$, όπου το K είναι το σύνολο όλων των λέξεων κλειδιά. Ο τύπος αυτός υποθέτει ανεξαρτησία ανάμεσα στις λέξεις, στις σελίδες και στις διαφημίσεις. Στη συνέχεια, αγνοεί τη σειρά και την εγγύτητα των όρων στη βαθμολογία. Ακολούθως, θα επικεντρωθούμε στον ορισμό του TaxScore. Η συνάρτηση αυτή δείχνει την τοπική αντιστοιχία ανάμεσα σε μια διαφήμιση και στη σελίδα. Εδώ, οι κλάσεις ταξινομούνται με μια ιεραρχία, σε αντίθεση με τις λέξεις κλειδιά που τις διαχειριζόμαστε σαν ανεξάρτητες διαστάσεις. Ένας από τους στόχους στην κατασκευή της συνάρτησης TaxScore είναι να μπορούμε να γενικεύουμε ανάμεσα στην ταξινόμηση, η οποία δέχεται τοπικά σχετικές μεταξύ τους διαφημίσεις. Η γενίκευση μπορεί να βοηθήσει στο να

τοποθετηθούν διαφημίσεις, σε περιπτώσεις που δεν υπάρχει ταίριασμα της διαφήμισης με την κατηγορία και τις λέξεις κλειδιά της σελίδας. Το παράδειγμα του σχήματος παρακάτω επιδεικνύει την διαδικασία.



Στο παράδειγμα αυτό, η απουσία διαφημίσεων για σκι, η σελίδα για σκι, η οποία περιέχει την λέξη «Atomic» μπορεί να συνδυαστεί με την διαθέσιμη διαφήμιση για snowboard της ίδιας εταιρίας.

Γενικά, είναι επιθυμητό το ταίριασμα να είναι μεγαλύτερο, όταν η διαφήμιση και η σελίδα κατατάσσονται στον ίδιο κόμβο, ενώ να είναι πιο ασθενής το ταίριασμα, όταν η απόσταση ανάμεσα στους κόμβους της κατάταξης γίνεται μεγαλύτερη. Υπάρχουν πολλαπλοί τρόποι για να ορίσουμε την απόσταση ανάμεσα σε δύο κόμβους κατάταξης. Δοσμένης μιας σελίδας, θα πρέπει να βρεθεί η διαφήμιση σε πολύ μικρό χρονικό διάστημα και αυτό γιατί η όποια καθυστέρηση έχει αντίκτυπο στον χρήστη. Έτσι έχουμε: TaxScore(PC,AC) = $\sum_{pc \in PC} \sum_{ac \in AC} idist(LCA(pc, ac), ac) cWeight(pc) cWeight(ac)$. Στη συνάρτηση αυτή, λαμβάνουμε υπόψη μας κάθε συνδυασμό τάξης σελίδας και τάξης διαφήμισης. Για κάθε συνδυασμό πολλαπλασιάζουμε το προϊόν των βαρών της τάξης με την αντίστροφη συνάρτηση idist(c1,c2). Η συνάρτηση αυτή παίρνει δύο κόμβους του ίδιου μονοπατιού και επιστρέφει έναν αριθμό από το διάστημα [0,1], ο οποίος εξαρτάται από την απόσταση των δύο τάξεων κόμβων. Επιστρέφει 1 αν οι δύο κόμβοι είναι οι ίδιοι και απορρίπτει προς το 0, όταν το LCA(pc,ac) ή αν το ac είναι η ρίζα της κατάταξης. Τέλος, προκειμένου να βρούμε το κόστος γενίκευσης από το παιδί στον πατέρα χρησιμοποιούμε απλές ευρετικές μεθόδους. Εδώ, κοιτάζουμε στο πεδίο εφαρμογής, όταν μετακινούμαστε από το παιδί στον πατέρα. Ο υπολογισμός του πεδίου γίνεται από την πυκνότητα των διαφημίσεων, οι οποίες κατατάσσονται στους κόμβους του πατέρα. Η πυκνότητα αποκτάται από την κατάταξη ενός μεγάλου συνόλου διαφημίσεων της κατάταξης. Βασιζόμενοι σε αυτό, ας ορίσουμε το n_c ως τον αριθμό των εγγράφων που έχουν καταταχθεί στο υποδέντρο με ρίζα το c. Στη συνέχεια, ορίζουμε : idist(c,p) = $\frac{n_c}{n_p}$, όπου το c αναπαριστά τον κόμβο του παιδιού και το p αναπαριστά τον κόμβο του γονέα.

5.1.4 Έρευνα στον χώρο των διαφημίσεων

Οι μεγαλύτερες k διαφημίσεις με το υψηλότερο σκορ προσφέρονται από το σύστημα για τοποθέτηση στη σελίδα του εκδότη. Η διαδικασία του υπολογισμού του σκορ και η επιλογή των διαφημίσεων πρέπει να γίνει σε πραγματικό χρόνο και έτσι θα πρέπει να είναι αποδοτικός. Οι συλλογές διαφημίσεων είναι στην εμβέλεια πολλών εισόδων και έτσι υπάρχει η ανάγκη για δεικτοδοτούμενη πρόσβαση στις διαφημίσεις. Για να είμαστε ικανοί να ψάχνουμε διαφημίσεις σε έναν συνδυασμό λέξεων κλειδιών και τάξεων, έχουμε αντιστοιχίσει την αντιστοιχία κατάταξης στην αντιστοιχία όρων και έχουμε υιοθετήσει τη συνάρτηση βαθμολόγησης να ταιριάζει και να υλοποιεί γρήγορη εκτίμηση στους ανεστραμμένους δείκτες. Χρησιμοποιήσαμε ένα πλαίσιο για ανεστραμμένους δείκτες, όπου υπάρχει μια λίστα για κάθε ξεχωριστό όρο. Οι διαφημίσεις περνιούνται στους όρους και κάθε όρος σχετίζεται με ένα βάρος το οποίο βασίζεται με το τμήμα στο οποίο εμφανίζεται. Τα βάρη ξεχωριστών εμφανίσεων ενός όρου σε μια διαφήμιση προστίθενται μαζί, έτσι ώστε οι λίστες να περιέχουν μια είσοδο για κάθε συνδυασμό όρου/διαφήμισης.

Το επόμενο στοίχημα είναι να καθοριστεί πώς θα κατατάξουμε τις διαφημίσεις, ώστε η πληροφορία της τάξης να διατηρείται στον δείκτη. Μια απλή μέθοδος είναι να δημιουργήσουμε μοναδικούς όρους για τις τάξεις και να σημειώσουμε κάθε διαφήμιση με έναν όρο για κάθε τάξη. Τα βάρη των όρων δίνονται από τη συνάρτηση idist() που περιγράφηκε παραπάνω. Όσον αφορά την πλευρά των αιτημάτων του χρήστη, με δεδομένα τις λέξεις κλειδιά και την τάξη της σελίδας, φτιάχνουμε ένα αίτημα μιας λέξης κλειδί, εισάγοντας τον όρο μιας κλάσης για κάθε πρόγονο τάξεων που έχουν αντιστοιχιστεί με τη σελίδα.

Η συνάρτηση βαθμολόγησης χρησιμοποιείται και για την τάξη των όρων, αλλά και για την τάξη των όρων κειμένου. Κατά τη διάρκεια του υπολογισμού της βαθμολογίας της τάξης, για κάθε μονοπάτι τάξης χρησιμοποιούμε μόνο τον όρο της μικρότερης τάξης, αγνοώντας τις άλλες. Για παράδειγμα, και από το παραπάνω διάγραμμα, αν μια διαφήμιση ανήκει στην τάξη του «Skiing» και σημειώνεται ταυτόχρονα με το «Skiing» και τον γονέα του «Winter Sports», ο δείκτης θα περιέχει την ειδική τάξη όρων για τα «Skiing» και «Winter Sports» με τα βάρη, σύμφωνα με το αποτέλεσμα της κατάταξης και της συνάρτησης idist. Όταν ταιριάζουμε με μια σελίδα, η οποία κατατάσσεται στο «Skiing», το αίτημα του χρήστη θα περιέχει όρους για την τάξη «Skiing» και για κάθε πρόγονο του. Όμως, όταν βαθμολογούμε μια διαφήμιση η οποία κατατάσσεται στο «Skiing» θα χρησιμοποιήσουμε το βάρος για τον όρο «Skiing». Οι διαφημίσεις οι οποίες κατατάσσονται στο «Snowboarding» θα βαθμολογηθούν με χρήση του βάρους του όρου «Winter Sports». Προκειμένου αυτός ο έλεγχος να γίνει αποτελεσματικά, διατηρούμε μια ταξινομημένη λίστα όλων των μονοπατιών των τάξεων και τη στιγμή της βαθμολόγησης, ερευνούμε τα μονοπάτια από το τέλος προς την κορυφή για κάποιον όρο που να εμφανίζεται στη διαφήμιση. Ο πρώτος όρος χρησιμοποιείται για βαθμολόγηση και οι υπόλοιποι αγνοούνται.

5.2 Μοντελοποίηση και ανάλυση για μη εγγυημένη παράδοση διαφήμισης (Post-click conversion modeling and analysis for non guaranteed delivery display advertising)

Η έννοια του display advertising, βασίζεται στο γεγονός ότι οι διαφημιστές πληρώνουν κάποιους εκδότες, ώστε να τοποθετήσουν τις διαφημίσεις τους στις ιστοσελίδες τους. Οι διαφημίσεις αυτές είναι γραφικής μορφής. Κάποιοι παράμετροι που σχετίζονται με τις διαφημίσεις αυτές, όπως το CTR (click-through rate) είναι παραδοσιακά ένα μέτρο της απόδοσης μια διαφήμισης. Επίσης, η παράμετρος CVR (conversion rate), μετρά την ποσότητα ορισμένων χρηστών που κάνουν μια προκαθορισμένη πράξη, όπως το να εγγραφούν κάπου, να κατεβάσουν ένα αρχείο κοκ. Τέλος, μια ακόμη παράμετρος που μας απασχολεί είναι το PCC (post-click conversion), δηλαδή η ανάλυση μιας εναλλακτικής διαδρομής μετά από το κλικ ενός χρήστη στην αναφερόμενη διαφήμιση. Κάποια στοιχεία τα οποία εξετάζουμε είναι η πιθανότητα εναλλακτικής (conversion) δοσμένου ενός κλικ στο περιεχόμενο μιας σελίδας, δηλαδή την πιθανότητα P(conversion|click,context).

5.2.1 Μοντέλα της παραμέτρου PPC (Post – click conversion)

Στο σημείο αυτό θα επικεντρωθούμε σε κάποια μοντέλα της παραμέτρου PPC, η οποία περιλαμβάνει διάφορους τύπους και ορισμούς χαρακτηριστικών, καθώς επίσης και το πρόβλημα της επιλογής χαρακτηριστικών.

5.2.1.1 Πρόβλεψη εναλλακτικής διαδρομής μετά από κλικ (Predicting conversions after a click)

Η διανομή χωρίς εγγύηση παράδοσης (NGD-non-guaranteed delivery) για τις διαφημίσεις είναι ένα μεγάλο και περίπλοκο υποσύστημα μιας γενικότερης αγοράς διαφημίσεων. Εδώ, οι εκδότες, οι διαφημιστές, τα δίκτυα διαφημίσεων, η ανταλλαγή διαφημίσεων, προσπαθούν να βελτιστοποιήσουν την τοποθέτηση μιας διαφήμισης με κάποιον τρόπο. Διάφορα στοιχεία είναι βασικά σε όλη αυτή τη διαδικασία, δηλαδή οι τιμές, οι εφευρέτες μιας διαφήμισης και η αναμενόμενη συμπεριφορά του χρήστη παίζουν το ρόλο τους. Θα επικεντρωθούμε στο σημείο αυτό, στον υπολογισμό μιας σημαντικής ποσότητας του συστήματος, η οποία είναι η πιθανότητα ένας χρήστης να προκαλέσει μια αντίθεση (conversion), αφότου κλικάρει σε μια συγκεκριμένη διαφήμιση. Είναι το επονομαζόμενο PPC (post-click conversion) πρόβλημα πρόβλεψης. Γίνεται χρήση των αρχείων από μία μεγάλη διαφημιστική πλατφόρμα προκειμένου να αναγνωρίσουμε τις αντιθέσεις και να τις αντιστοιχίσουμε με τα διαφημιστικά κλικ. Πιο συγκεκριμένα, ενδιαφερόμαστε να μοντελοποιήσουμε την υπό συνθήκη πιθανότητα ότι ο χρήστης έκανε κλικ σε μια σχετική διαφήμιση. Δύο κατηγορίες γεγονότων θα μας απασχολήσουν για να προσεγγίσουμε το πρόβλημα: 1) γεγονότα, όπου ο χρήστης κάνει κλικ σε μια διαφήμιση, αλλά δεν μετατρέπονται και θεωρούνται αρνητικές περιπτώσεις μετατροπής 2) γεγονότα, όπου ο χρήστης κάνει κλικ σε μια διαφήμιση και μετατρέπονται και έτσι θεωρούνται θετικές περιπτώσεις μετατροπής.

Ένα σύνολο γεγονότων αντιπροσωπεύεται από το $D = \{x(bi,ai,ui), yi\}$, όπου το $x(bi,ai,ui) \in R$ είναι μια αναπαράσταση της διαθέσιμης πληροφορίας για ένα γεγονός, συνήθως συσχετιζόμενο με τον εκδότη bi, την διαφήμιση ai και τον χρήστη ui. Το $yi \in \{0,+1\}$ είναι

ένας δείκτης για το αν μια μετατροπή συνδέεται με ένα γεγονός. Έτσι, όταν δίνεται μερική ή όλη η πληροφορία ενός γεγονότος, σκοπός είναι να παραχθεί η πιθανότητα p(y|b,a,u) ή πιο απλά, η πιθανότητα p(y|x). Ουσιαστικά, θα χρησιμοποιήσουμε ένα γραμμικό μοντέλο, που θα συνδυάζει την μεμονωμένη συνεισφορά κάθε στοιχείου του διανύσματος χαρακτηριστικών διαμέσου του βάρους w = (w1,…,wd). Πιο συγκεκριμένα, p(y=1|b,a,u) = $\frac{1}{1+\exp\left(-\sum_{i=1}^{d} wixi(b,a,u)\right)}$, όπου το wi ∈ R και σταθμίζει τη συνεισφορά των χαρακτηριστικών. Τελικά, βρίσκουμε τις βέλτιστες τιμές για το wi με χρήση του κριτηρίου της πιθανότητας w* = argmaxp(w|D). Θα θέσουμε το p(w) = N(w;0,Σ) στο w και έχουμε: w* = $\text{argmax}_{w\in R}{}^{d}$ $\sum_{i=1}^{N} \log\big(p(yi|bi,ai,ui)\big) + logp(w)$.

5.2.2 Χαρακτηριστικά

Όταν χτίζουμε τέτοιου είδους μοντέλα είναι σημαντικό να κρατείται πληροφορία για τις διαφορετικές πτυχές μιας προσωπικότητας. Γενικά, αναφερόμαστε σε τέσσερα σύνολα χαρακτηριστικών, τα οποία λαμβάνουμε υπόψη μας. Είναι χαρακτηριστικά που αφορούν πληροφορία που αποκτάται τη στιγμή που ο χρήστης κλικάρει σε μια διαφήμιση μιας ιστοσελίδας. Κάποια από αυτά λοιπόν, συνοψίζονται στον πίνακα που παρουσιάζεται στη συνέχεια. Ενδεικτικά με τη μελέτη των χαρακτηριστικών αυτών γίνεται η απλή παρατήρηση ότι οι διαφημίσεις που δουλεύουν καλύτερα στο παρελθόν, πιθανόν να δουλέψουν το ίδιο καλά και μελλοντικά.

Feature family	Feature members
Advertiser	advertiser (id), advertiser network, campaign, creative, conversion id, ad group, ad size, creative type, offer type
Publisher	publisher (id), publisher network, site, section, url, page referrer
User (when avail.)	gender, age, region, network speed, accept cookies, geo, user segments
Other	serve time

Στο μοντέλο τύπου PCC, τα στατιστικά, όπως είναι ο συνολικός αριθμός των κλικς και ο ρυθμός μετατροπής χρησιμοποιούνται συχνά για τη μέτρηση της λειτουργίας των διαφημίσεων. Επίσης, προκειμένου να μετρηθεί η ποικιλία μιας λειτουργίας διαφήμισης για διαφορετικούς εκδότες, παρόμοια χαρακτηριστικά μπορούν να αποκτηθούν με το να ζευγαρώνονται διαφημιστές σε διαφορετικά επίπεδα με τους εκδότες. Για παράδειγμα, το ζευγάρι publisher-creative ή id-creative. Παρόμοια, ομαδοποιήσεις όπως η ηλικία, το φύλο μαζί με το αντικείμενο διαφήμισης (creative) είναι καλά εφόδια, επειδή μπορούν να κρατήσουν την ποικιλία του παρελθόντος όσον αφορά τις διαφημίσεις για διαφορετικούς χρήστες.

5.2.3 Αυτοματοποιημένη ανάλυση χαρακτηριστικών (automated feature analysis)

Παρατηρούμε ότι η τιμή ενός χαρακτηριστικού, η οποία λαμβάνει χώρα με κάποια συχνότητα στα δεδομένα δοκιμής, δεν θα συμβεί με τον ίδιο τρόπο απαραίτητα και on-line. Πιο συγκεκριμένα, το I(Xj,Y) αναπαριστά την πληροφορία του χαρακτηριστικού Xj για το Y. Μια ευρέως χρησιμοποιούμενη μέθοδος επιλογής χαρακτηριστικών αποτελείται από την

επιλογή του χαρακτηριστικού Xj ώστε το I(Xj,Y) να είναι υψηλό, με το Y να είναι το χαρακτηριστικό στόχου. Για παράδειγμα, θα μπορούσαν να επιλεχθούν τα μεγαλύτερα σε βαθμολογία χαρακτηριστικά. Για χαρακτηριστικά με δυναμική συνεισφορά, χρησιμοποιούμε την Xu σαν μια τυχαία μεταβλητή, η οποία παίρνει μοναδικές τιμές. Έτσι, το I(Xu,Y) = H(Y), αφού οι τιμές του Xu μπορούν να αναγνωρίσουν πλήρως το σημείο της πληροφορίας και κατ' επέκταση την ετικέτα. Οπότε έχουμε: I(Xu,Y) = $\sum_{x,y} p(x_u, y) log p(y|x_u)/p(y)$ = H(Y), όπου το p(y*|xu)=1, για κάποια y=y*. Παρόλα αυτά, το Xu δεν είναι χρήσιμο σαν χαρακτηριστικό για την πρόβλεψη του y, αφού οι τιμές του είναι μοναδικές και δεν παρατηρούνται σε κάποιο σύνολο δοκιμών. Προκειμένου να αναφερθούμε στο πρόβλημα αυτό, χρησιμοποιούμε μια σχετική συνάρτηση που αφορά μια κατανομή αναφοράς. Αυτή δίνεται από το $\tilde{p}$(x, y), και στη συνέχεια ορίζουμε το MI σε σχέση την κατανομή αναφοράς ως: $I_p(Xi, Y) = \sum_{xi,y} \tilde{p}(xi, y) log \frac{p(xi,y)}{p(xi)p(y)}$. Ο ορισμός αυτός παρουσιάζει το πρόβλημα ότι η συχνότητα log δεν είναι καλά ορισμένη στις περιπτώσεις που p(xi) = 0. Αυτό συμβαίνει, όταν η τιμή ενός χαρακτηριστικού την έχουμε συναντήσει στην κατανομή αναφοράς $\tilde{p}$, αλλά όχι στο σύνολο δοκιμών της κατανομής p. Έτσι, χρησιμοποιούμε μια κανονικοποίηση του συνόλου δοκιμών της μορφής: $p_r(xi, y) = \frac{N_p(xi,y)+p(y)}{N+|Xi|}$, με $p_r(xi)$> 0 και όπου το |Xi| είναι ο αριθμός των καταστάσεων που έχουν καταληφθεί από το Xi. Είναι πιθανόν να δείξουμε ότι αν $(\forall y)p(y) > 0$ αυτό δεν επηρεάζει την κατανομή στόχου, όπου $p_r(y)=p(y)$. Σε τελευταία περίπτωση, έχουμε ότι η συχνότητα log γίνεται 0. Η κύρια σχετική ιδιότητα της καινούργιας ποσότητας πληροφορίας είναι ότι αφού τα χαρακτηριστικά εκτιμούνται σε μια κατανομή αναφοράς, κάποιες σχέσεις τις οποίες συναντάμε σε ένα συγκεκριμένο σύνολο πληροφορίας και δεν γενικεύουμε στο σύνολο δοκιμών, συχνά αγνοούνται.

Ας σκεφτούμε το ίδιο ενδεικτικό παράδειγμα όπως παραπάνω: οι τιμές του Xu που δεν εμφανίζονται στην κατανομή δοκιμής, και έτσι η κατανομή δοκιμής δε θα επιβαρύνει με επιπλέον όγκο τις τιμές αυτές και η Xu δε θα έχει πληροφορία για τη συνάρτηση στόχου που μας ενδιαφέρει. Πιο επίσημα, $I_p(Xu,Y) = \sum_{x_u,y} \tilde{p}(x_u, y) log \frac{p_r(x_u,y)}{p_r(x_u)p_r(y)}$ = 0. Να σημειωθεί ότι σε πολλές περιπτώσεις, οποιαδήποτε κατανομή αναφοράς $\tilde{p}$, υπολογίζεται σε ένα έγκυρο δείγμα, διαφορετικό από την κατανομή δοκιμής θα επιτρέψει την προτεινόμενη μέτρηση, ώστε να αποφευχθούν σχέσεις που αφορούν το σύνολο δοκιμών.

5.3 Κίνδυνος μεγιστοποίησης εσόδων στην κατηγορία display advertising (Risk aware revenue maximization in display advertising)

Η διαφήμιση της κατηγορίας display advertising είναι η γραφική διαφήμιση του παγκόσμιου ιστού (WWW), που εμφανίζεται δίπλα από το περιεχόμενο των ιστοσελίδων, των e-mail κοκ. Την προηγούμενη δεκαετία, αυτού του είδους οι διαφημίσεις έχουν εξελιχθεί από απλές διαφημίσεις τύπου banner και pop-up, σε διάφορους συνδυασμούς κειμένου, εικόνας, ήχου, βίντεο κλπ. Όσον αφορά την αγορά, έχει αποδειχθεί ότι το display advertising συνεχίζει να δείχνει σημαντική ανάπτυξη, όπως αποδεικνύουν εταιρίες, όπως η Microsoft, η Yahoo, η Google. Σαν μια διαδικασία πώλησης, οι διαφημίσεις πωλούνται σαν πακέτα, και είναι το αποτέλεσμα διαπραγμάτευσης ανάμεσα σε πωλήσεις και σε

διαφημιστικούς πράκτορες. Στο σημείο αυτό θα ασχοληθούμε με το πρόβλημα της μεγιστοποίησης των εσόδων προσδίδοντας τιμές σε αυτά που παρουσιάζονται στην οθόνη. Χειριζόμαστε τη διαδικασία αυτή σαν ένα πρόβλημα κατανομής. Σκοπός λοιπόν, είναι η μεγιστοποίηση του κέρδους και η επαναλαμβανόμενη προσαρμογή των τιμών. Ο κύριος άγνωστος για την μεγιστοποίηση των εσόδων στο περιβάλλον του display advertising είναι το πώς η ζήτηση για τις διαφημίσεις αλλάζει τις τιμές, η κλασική καμπύλη ζήτησης.

5.3.1 Διατύπωση του προβλήματος και παραδοχές

Στο σημείο αυτό θα εισάγουμε το μαθηματικό μοντέλο, το οποίο αποτυπώνει την διαδικασία πωλήσεων, εισάγοντας ένα μοντέλο μετρήσεων, τις παραμέτρους και τις υποθέσεις που χρησιμοποιούνται στην ανάλυση. Μια ακολουθία από διαφημιστές θα φτάσει σε έναν εκδότη, σύμφωνα με κάποια διαδικασία στο χρόνο. Ο διαφημιστής α, $\alpha \in A$, φέρνει το χρηματικό ποσό (budget) Βα, το οποίο είναι διαθέσιμο για να δαπανηθεί σε ένα σύνολο εμφανίσεων. Στη συνέχεια, οι διαφημιστές συχνά έχουν κάποιες αρχικές προτιμήσεις σε σχέση με είδη διαφημίσεων που θέλουν να αγοράσουν, αλλά και συγκεκριμένες προσδοκίες λειτουργικότητας, οι οποίες δεν έχουν καταγραφεί πρακτικά, αλλά επηρεάζουν τη δυναμική διαπραγμάτευση. Υποθέτουμε ότι τα χρηματικά ποσά διαπραγμάτευσης είναι ανεξάρτητα από τα σημεία αφίξεων στο χρόνο. Η ανάλυση της πραγματικής κίνησης διαπραγμάτευσης υπονοεί ότι ο ρυθμός άφιξης ενός διαφημιστή είναι μικρότερος από τον ρυθμό άφιξης όλων των διαφημιστών.

Από την άλλη πλευρά, υποθέτοντας ότι οι αγοραστικές αποφάσεις διαφορετικών διαφημιστών είναι αμοιβαία ανεξάρτητες, διαχειριζόμαστε ένα σύνολο από συναλλαγές που συμβαίνουν σε μια ορισμένη χρονική περίοδο (φάση πειραμάτων) σαν να είναι και αυτή ανεξάρτητη. Κάθε διαφημιστής έχει ένα μοναδικό πράκτορα πωλήσεων το οποίο διαχειρίζεται όλες τις καμπάνιες. Ένας πράκτορας ξεκινά μια διαπραγμάτευση, προσπαθώντας να διεκπεραιώσει όλους τους στόχους της διαφημιστικής καμπάνιας, δεσμεύοντας ένα επιθυμητό πακέτο εμφανίσεων, το οποίο συμβολίζεται εδώ από ένα K-διαστάσεων διάνυσμα $(M_{a1}(\tilde{p}),\ldots,M_{aK}(\tilde{p}))$. Εκτός από τους στόχους της καμπάνιας, η απόφαση του πράκτορα πωλήσεων εξαρτάται από τα πρόσφατα επίπεδα χρήσης και τιμών. Δυστυχώς, δεν κρατούνται καταγραφές από τη διαδικασία διαπραγμάτευσης, εκτός από το τελικό πακέτο των συμφωνημένων εμφανίσεων. Οι δεσμεύσεις εμφανίσεων είναι μια συνάρτηση του τρέχοντος διανύσματος τιμών $\tilde{p} = (p_1,p_2,\ldots p_k,)$, όπου το p_k , $1 \le k \le K$, είναι η τιμή της καταγραφής του k. Χρησιμοποιούμε το $U_k(\tilde{p})$, $1 \le k \le K$, για να συμβολίσουμε το ποσοστό καταγραφής k, το οποίο χρησιμοποιείται στην τρέχουσα χρονική στιγμή κατά τη διάρκεια της συναλλαγής. Το πακέτο εμφανίσεων της διαπραγμάτευσης $(M_{a1}(\tilde{p}),\ldots,M_{aK}(\tilde{p}))$, χρειάζεται να προστεθεί στο διαθέσιμο χρηματικό ποσό το οποίο δαπανάται, για παράδειγμα $Ba = p_1M_{a1}(\tilde{p}) + \ldots + p_kM_{ak}(\tilde{p})$. Για να συνοψίσουμε, η καταγραφή πληροφορίας από παλαιότερες πωλήσεις /συναλλαγές περιλαμβάνει: α) το όνομα του διαφημιστή, ας πούμε α και τον πράκτορα πώλησης β) το χρηματικό ποσό Βα γ) την διαθεσιμότητα γονέα τη στιγμή της πώλησης, $(1\text{-}U_1(\tilde{p}),\ldots,1\text{-}U_k(\tilde{p}))$ και δ) το τελικό πακέτο των εμφανίσεων διαπραγμάτευσης $(M_{a1}(\tilde{p}),\ldots,M_{aK}(\tilde{p}))$.

Δεν κάνουμε υποθέσεις για τη διαδικασία άφιξης των διαφημιστών, των χρημάτων, τη διάρκεια του συμβολαίου. Παρόλα αυτά, υποθέτουμε επαρκή σταθερότητα για

υπολογισμούς από μια χρονική περίοδο, προκειμένου να διατηρούμαστε σταθεροί σε επόμενες χρονικές περιόδους. Ο σκοπός του κάθε εκδότη είναι να αυξάνει τα έσοδα που υπόκεινται στη διαθέσιμη απογραφή. Τα συνολικά έσοδα ισούνται με το άθροισμα των εσόδων που έχουν δαπανηθεί για όλες τις απογραφές, για παράδειγμα $R(\bar{p}) = \sum_{a\in A} Ba = \sum_{a\in A}\sum_{k=1}^{K} M_{ak}(\bar{p})p_k$. Επομένως, το πρόβλημα βελτιστοποίησης που αντιμετωπίζει ο εκδότης είναι το εξής: $max_{\bar{p}} \sum_{k=1}^{K} p_k \sum_a M_{ak}(\bar{p})$, $\sum_a M_{ak}(\bar{p}) \le C_k$, k=1,…,K.

5.3.2 Βελτιστοποιημένος αλγόριθμος

Στην παράγραφο αυτή, προτείνουμε έναν βελτιστοποιημένο αλγόριθμο και δείχνουμε ότι κάθε προτεινόμενη ανανέωση τιμής αυξάνει τα αναμενόμενα έσοδα του εκδότη. Ας σημειωθεί επίσης, ότι η χρήση του k σαν άνω όριο, $U_k(\bar{p})$ μπορεί να επανεκφραστεί και σαν το άθροισμα των κατανομών σε όλους τους διαφημιστές, $C_k\ U_k(\bar{p}) = \sum_a M_{ak}(\bar{p})$. Ο αλγόριθμος σε βήματα αποτελείται από την επανάληψη των τριών ακόλουθων βημάτων:

1. Υπολογισμός του ανάδελτα[2] (gradient) $\nabla L(\bar{\lambda}, \bar{p})$ στην τρέχον σημείο τιμής, όπου το $\lambda_k = \lambda_k(\bar{p}) = p_k\ U_k(\bar{p})$. Για παράδειγμα, υποθέτουμε ότι το κόστος της καταγραφής που δεν έχει πωληθεί, ισούται με την πραγματική απογραφή της τιμής της μονάδας.
2. Προσαρμογή της τιμής διανύσματος στην κατεύθυνση του υπολογισμένου ανάδελτα.
3. Επανέλαβε τα βήματα 1 και 2 μέχρι το $\nabla L(\bar{\lambda}, \bar{p}) = 0$.

Θεώρημα 1

Υποθέτουμε ότι αν τα ανάδελτα έχουν υπολογιστεί με ακρίβεια, τότε ο παραπάνω αλγόριθμος μεγιστοποιεί τα έσοδα σε κάθε επανάληψη και συγκλίνει στο τοπικό μέγιστο της συνάρτησης εσόδων.

Η απόδειξη του παραπάνω θεωρήματος και ο υπολογισμός του ανάδελτα μπορεί να βρεθεί στη σχετική βιβλιογραφία της εργασίας.[3]

Βιβλιογραφία κεφαλαίου

[1]Andrei Broder, Marcus Fontoura, Vanja Josifovski, Lance Riedel. “A Semantic Approach to Contextual Advertising”. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval.* New York City, USA. 2007.

[2] Romer Rosales, Haibin Cheng, Eren Manavoglu. “Post-Click Conversion Modeling and Analysis for Non-Guaranteed Delivery Display Advertising”. In *Proceedings of the Fifth International Conference on Web Search and Web Data Mining, WSDM 2012,* Seattle, WA, USA, February 8-12, 2012.

[2] Είναι διανυσματικός διαφορικός τελεστής των μερικών παραγώγων μιας συνάρτησης ως προς τις τρεις διαστάσεις του χώρου.

[3] William D. Heavlin, Ana Radovanovic. “Risk – Aware Revenue Maximization in Display Advertising”. In *Proceedings of the 21st international conference on World Wide Web,* Lyon, France, April 16-20, 2012.

Συνολική βιβλιογραφία

[1] Panagiotis Papadimitriou. “Algorithms and Strategies for Web Advertising”. Stanford University, Dept. of Electrical Engineering. December 2011

[2] Abhirup Nath, Shibnath Mukherjee, Prateek Jain, Navin Goyal, Srivatsan Laxman. “Ad Impression Forecasting for Sponsored Search”. In *Proceedings of the 22nd international conference on World Wide Web (WWW 2013).* Rio de Janeiro, Brazil. May 13-17, 2013.

[3] Anand Rajaraman, Jure Leskovec, Jeffrey D. Ullman. “Mining of Massive Datasets” (Chapter 8). Stanford University. 2013.

[4] Sujith Ravi, Andrei Broder, Engeniy Gabrilovich, Vanja Josifovski, Sandeep Pandey, Bo Pang. “Automatic Generation of Bid Phrases for Online Advertising”. In *Proceedings of the third ACM international conference on Web search and data mining, Pages 341-350 (WSDM 2010).* New York City, USA. February 3-6, 2010.

[5] Chinmay Karande, Aranyak Mehta, Ramakrishnan Srikant. “Optimizing Budget Constrained Spend in Search Advertising”. In *Proceedings of the sixth ACM international conference on Web search and data mining, Pages 697-706 (WSDM 2013).* Rome, Italy. February 4-8, 2013.

[6] Michael Bendersky, Evgeniy Gabrilovich, Vanja Josifovski, Donald Metzler. “The anatomy of an Ad: Structured Indexing and Retrieval for Sponsored Search”. In *Proceedings of the 19th international conference on the World Wide Web, (WWW 2010)*. Raleigh, North Carolina, USA. April 26-30, 2010.

[7] Yejin Choi, Marcus Fontoura, Evgeniy Gabrilovich, Vanja Josifovski, Mauricio Mediano, Bo Pang. “Using Landing Pages for Sponsored Search Ad Selection”. In *Proceedings of the 19th international conference on the World Wide Web, (WWW 2010)*. Raleigh, North Carolina, USA. April 26-30, 2010.

[8] Zeyuan Allen Zhu, Weizhu Chen, Tom Minka, Chenguang Zhu, Zheng Chen. “A Novel Click Model and Its Applications to Online Advertising”. In *Proceedings of the third ACM international conference on Web search and data mining, (WSDM 2010).* New York City, USA. February 4-6, 2010.

[9] Eric Sodomka, Sebastien Lahaie, Dustin Hillard. “A Hierarchical Model for Value Estimation in Sponsored Search”. In *EC 2011 Seventh Ad Auctions Workshop*. June 2011.

[10] Abhirup Nath, Shibnath Mukherjee, Prateek Jain, Navin Goyal, Srivatsan Laxman. “Ad Impression Forecasting for Sponsored Search”. In *Proceedings of the 22nd international conference on World Wide Web Pages 943-952(WWW 2013).* Rio de Janeiro, Brazil. May 13-17, 2013.

[11] Xingxing Wang, Shijie Lin, Dongying Kong, Liheng Xu, Qiang Yan, Siwei Lai, Liang Wu, Alvin Chin, Guibo Zhu, Heng Gao, Yang Wu, Danny Bickson, Yuanfeng Du, Neng Gong,

Chengchun Shu, Shuang Wang, Kang Liu, Shuren Li, Jun Zhao, Fei Tan, Yuanchun Zhou. “Click-Through Prediction for Sponsored Search Advertising with Hybrid Models”. *ACM KDD CUP Workshop*, 2012.

[12] Nikolay Archak, Vahab S. Mirrokni, S. Muthukrishnan. “Mining Advertiser-specific User Behavior Using Adfactors”. In *Proceedings of the 19th international conference on the World Wide Web, Pages 41-40(WWW 2010)*. Raleigh, North Carolina, USA. April 26-30, 2010.

[13] Andrei Broder, Marcus Fontoura, Vanja Josifovski, Lance Riedel. “A Semantic Approach to Contextual Advertising”. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval.* New York City, USA. 2007.

[14] Romer Rosales, Haibin Cheng, Eren Manavoglu. “Post-Click Conversion Modeling and Analysis for Non-Guaranteed Delivery Display Advertising”. In *Proceedings of the Fifth International Conference on Web Search and Web Data Mining, WSDM 2012,* Seattle, WA, USA, February 8-12, 2012.

[15] William D. Heavlin, Ana Radovanovic. “Risk – Aware Revenue Maximization in Display Advertising”. In *Proceedings of the 21st international conference on World Wide Web,* Lyon, France, April 16-20, 2012.

[16] http://en.wikipedia.org/wiki/Online_advertising

[17] http://www.techopedia.com/definition/26362/online-advertising

[18] http://www.entrepreneur.com/encyclopedia/online-advertising

[19] http://en.wikipedia.org/wiki/Web_banner