Εξόρυξη δεδομένων και συμπεριφορά χρήστη στον Ιστό

μια βιβλιογραφική μελέτη στο μάθημα Εξόρυξη δεδομένων και Αλγόριθμοι Μάθησης

Άγκο Μπεσιάνα 1059662
Ζεκυριά Αθανασία 1059660

Διδάσκοντας: Μακρής Χρήστος

Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής
Πανεπιστήμιο Πατρών

Ελλάδα
2021

Περιεχόμενα

Περιεχόμενα 2

Εισαγωγή 3

1. Εξόρυξη δεδομένων για εξατομίκευση στο διαδίκτυο (Web mining for Web personalization) 5

1.1 Βασικές πληροφορίες 5

1.2 Εξατομίκευση Ιστού 6

1.3 Δημιουργία προφίλ χρήστη 7

1.4 Συλλογή δεδομένων 7

1.5 Ανάλυση αρχείων καταγραφής και εξόρυξη στοιχείων χρήσης ιστού 8

1.6 Ερευνητικές Πρωτοβουλίες 9

2. Μοντέλο πρόβλεψης ιστοσελίδων που βασίζεται στην αναπαράσταση δέντρου click-stream της συμπεριφοράς του χρήστη (A Web Page Prediction Model Based on Click-Stream Tree Representation of User Behavior) 11

2.1 Πρόβλεψη συμπεριφοράς χρήστη 11

2.2.1. Προετοιμασία δεδομένων και καθαρισμός 12

2.2.2. Μέτρηση ομοιότητας περιόδου λειτουργίας 12

2.2.3 Συσταδοποίηση ανά ζεύγη 13

2.2.4 Αναπαράσταση συστάδων 13

2.2.5 Μηχανή προτάσεων 14

2.3 Αποτελέσματα πειράματος 15

3. Έγκαιρος εντοπισμός εξόδων χρήστη από δεδομένα click-stream : Ένα Markov Μοντέλο Διεργασίας Με Διαμόρφωση Μαρκοειδών Σημείων (Early Detection of User Exits from Clickstream Data: A Markov Modulated Marked Point Process Model) 16

4. HPM:Υβριδικό μοντέλο για την πρόβλεψη της συμπεριφοράς του χρήστη με βάση την ανάλυση N-Gram και τα αρχεία καταγραφής πρόσβασης (HPM: A Hybrid Model for User's Behavior Prediction Based on N-Gram Parsing and Access Logs) 21

Βιβλιογραφία 25

Εισαγωγή

Η πρόβλεψη της πρόθεσης των χρηστών για ένα συγκεκριμένο προϊόν ή κατηγορία προϊόντων, με βάση τις αλληλεπιδράσεις μέσα σε έναν ιστότοπο, είναι κρίσιμη για τοποθεσίες ηλεκτρονικού εμπορίου και δίκτυα προβολής διαφημίσεων, ιδιαίτερα για τον επαναπροσδιορισμό στόχων. Παρακολουθώντας τα μοτίβα αναζήτησης των καταναλωτών, οι διαδικτυακοί έμποροι μπορούν να κατανοήσουν καλύτερα τη συμπεριφορά και τις προθέσεις τους [1].
Στο ηλεκτρονικό εμπόριο μέσω φορητών συσκευών και κινητών τηλεφώνων υπάρχει διαθέσιμο ένα πλούσιο σύνολο δεδομένων και οι πιθανοί καταναλωτές αναζητούν πληροφορίες για τα προϊόντα πριν από τη λήψη αγοραστικών αποφάσεων, αντικατοπτρίζοντας έτσι τις προθέσεις των καταναλωτών για αγορές. Οι χρήστες εμφανίζουν διαφορετικά μοτίβα αναζήτησης, δηλαδή τον χρόνο που δαπανάται ανά στοιχείο, τη συχνότητα αναζήτησης και τις επισκέψεις που επιστρέφουν [2]. Τα δεδομένα Click-stream μπορούν να χρησιμοποιηθούν για τον ποσοτικό προσδιορισμό της συμπεριφοράς αναζήτησης με τη χρήση τεχνικών μηχανικής μάθησης, που επικεντρώνονται κυρίως σε αρχεία αγορών. Ενώ η αγορά υποδεικνύει στους καταναλωτές τις τελικές προτιμήσεις στην ίδια κατηγορία, η αναζήτηση αποτελεί επίσης ένα σημαντικό στοιχείο για τον προσδιορισμό πρόθεσης προς μια συγκεκριμένη κατηγορία [3].

Στην παρούσα εργασία, εμβαθύνουμε σε τέσσερα επιστημονικά άρθρα με κοινό γνώμονα την πρόβλεψη συμπεριφοράς χρήστη στον Ιστό. Πιο συγκεκριμένα,

1. Το πρώτο άρθρο [4] πραγματεύεται την **εξατομίκευση του Ιστού**. Μια διαδικασία προσαρμογής μιας τοποθεσίας Ιστού στις ανάγκες συγκεκριμένων χρηστών, με την αξιοποίηση των γνώσεων που αποκτώνται από την ανάλυση της συμπεριφοράς πλοήγησης του χρήστη (δεδομένα χρήσης) σε συνδυασμό με άλλες πληροφορίες που συλλέγονται στο πλαίσιο του Ιστού, δηλαδή τη δομή, το περιεχόμενο, και τα δεδομένα προφίλ του χρήστη. Σε αυτό το άρθρο παρουσιάζεται μια έρευνα σχετικά με τη χρήση της εξόρυξης Ιστού για εξατομίκευση του Ιστού. Πιο συγκεκριμένα, θα παρουσιάσουμε τις υπομονάδες που αποτελούν ένα σύστημα εξατομίκευσης του Ιστού, δίνοντας έμφαση στη λειτουργική μονάδα εξόρυξης στοιχείων χρήσης του Ιστού. Παρουσιάζεται μια επισκόπηση των πιο συνηθισμένων μεθόδων που χρησιμοποιούνται, καθώς και τεχνικών ζητημάτων που προκύπτουν, μαζί με μια σύντομη επισκόπηση των πιο δημοφιλών εργαλείων και εφαρμογών που διατίθενται από προμηθευτές λογισμικού. Επιπλέον, παρουσιάζονται οι πιο σημαντικές ερευνητικές πρωτοβουλίες στους τομείς εξόρυξης και εξατομίκευσης της χρήσης του Διαδικτύου.

2. Το δεύτερο επιστημονικό άρθρο [5] ασχολείται με ένα **μοντέλο πρόβλεψης ιστοσελίδας**. Η πρόβλεψη της επόμενης αίτησης ενός χρήστη κατά την επίσκεψή του σε ιστοσελίδες αποκτά πολύ μεγάλη σημασία καθώς αυξάνεται η δραστηριότητα στον Ιστό. Τα μοντέλα Markov και οι παραλλαγές τους, ή τα μοντέλα τους που βασίζονται στην εξόρυξη ακολουθίας έχουν βρεθεί κατάλληλα για αυτό το πρόβλημα. Ωστόσο, τα μοντέλα Markov ανώτερης τάξης είναι εξαιρετικά περίπλοκα λόγω του μεγάλου αριθμού των καταστάσεων τους, ενώ τα μοντέλα Markov κατώτερης τάξης

δεν αποτυπώνουν ολόκληρη τη συμπεριφορά ενός χρήστη σε μια συνεδρία. Τα μοντέλα που βασίζονται στην εξόρυξη σειριακών σχημάτων λαμβάνουν υπόψη μόνο τις συχνές ακολουθίες στο σύνολο δεδομένων, καθιστώντας δύσκολο να προβλεφθεί το επόμενο αίτημα μετά από μια σελίδα που δεν βρίσκεται στο σειριακό μοτίβο. Επιπλέον, είναι δύσκολο να βρεθούν μοντέλα για την εξόρυξη δύο διαφορετικών ειδών πληροφοριών από μια περίοδο εργασίας χρήστη. Προτείνεται, λοιπόν, ένα νέο μοντέλο που λαμβάνει υπόψη, τόσο τις πληροφορίες σειράς των σελίδων σε μια συνεδρία, όσο και τον χρόνο που δαπανάται σε αυτές. Οι περίοδοι χρήσης ομαδοποιούνται με βάση την ομοιότητά τους ως προς ζεύγη και αναπαριστούν τις συστάδες που προκύπτουν από ένα δέντρο click-stream. Στη συνέχεια, η νέα περίοδος λειτουργίας χρήστη αντιστοιχίζεται σε ένα σύμπλεγμα με βάση ένα μέτρο ομοιότητας. Το δέντρο click-stream αυτού του συμπλέγματος χρησιμοποιείται για τη δημιουργία του συνόλου προτάσεων. Το μοντέλο μπορεί να χρησιμοποιηθεί ως μέρος ενός συστήματος προανάκτησης κρυφής μνήμης καθώς και ως μοντέλο προτάσεων.

3. Το τρίτο άρθρο [6] ασχολείται επίσης με πρόβλεψη και πιο συγκεκριμένα, με τον **εντοπισμό εξόδων χρήστη**, εκ νέου μέσω click-stream δεδομένων και με μοντέλα Markov. Ένας αρκετά μεγάλος αριθμός χρηστών εξέρχεται από ιστοσελίδες εμπορίου χωρίς να προμηθευτεί κάποιο προϊόν, για το λόγο αυτό δημιουργήθηκε ένα μοντέλο κατάλληλα διαμορφωμένο για την ανίχνευση αυτών των χρηστών. Το συγκεκριμένο μοντέλο μας αποτυπώνει μια ακολουθία από επισκέψεις του χρήστη στη σελίδα αλλά και το χρόνο που δαπανήθηκε. Εκτός από το μοντέλο αυτό παρουσιάστηκε ένα πλαίσιο αξιολόγησης του κινδύνου εξόδου χωρίς αγορά χρησιμοποιώντας μια βαθμολογία κινδύνου. Χρησιμοποιώντας τα προτεινόμενα μοντέλα οι ιδιοκτήτες ιστότοπων ηλεκτονικού εμπορίου μπορούν να εστιάσουν καλύτερα την προσοχή τους στους χρήστες και να παρεμβαίνουν εγκαίρως. Επίσης, πειράματα έδειξαν ότι τα μοντέλα αυτά είναι περισσότερο αποτελεσματικά διότι ανιχνεύουν τους χρήστες με μεγαλύτερη ακρίβεια και 78,4 % νωρίτερα.

4. Τέλος, το τέταρτο [7] περιγράφει ένα υβριδικό **μοντέλο πρόβλεψης συμπεριφοράς χρήστη**. Η πρόβλεψη της συμπεριφοράς των χρηστών στον Παγκόσμιο Ιστό υπήρξε κρίσιμο ζήτημα τα τελευταία χρόνια. Για το λόγο αυτό δημιουργήθηκε ένα υβριδικό μοντέλο πρόβλεψης το οποίο αξιοποιεί όλες τις πληροφορίες χρήσης του web από τους χρήστες καθώς και του περιεχομένου των ιστοσελίδων. Συγκεκριμένα, το μοντέλο αυτό χρησιμοποιεί Query-URL click-graph χρήση ερωτημάτων που υποβλήθηκαν σε χρήστες με αντίστοιχες διευθύνσεις URL και χωρίζεται σε 2 φάσεις (offline και online). Η offline περιλαμβάνει τη χρήση δεδομένων από αρχεία καταγραφής πρόσβασης για να οδηγήσει στην πρόβλεψη της συμπεριφοράς των χρηστών, ενώ η online σχετίζεται με την αλληλεπίδραση του χρήστη με το σύστημα. Τέλος η αποτελεσματικότητα του μοντέλου αυτού αποδείχτηκε πειραματικά παρόλα αυτά είναι απαραίτητη η βελτίωση του για να ταιριάζει επακριβώς με τους όρους του ερωτήματος του ενδιαφέροντος του χρήστη με τα αρχεία καταγραφής.

1. Εξόρυξη δεδομένων για εξατομίκευση στο διαδίκτυο (Web mining for Web personalization)

1.1 Βασικές πληροφορίες

Τα δεδομένα Ιστού είναι εκείνα που μπορούν να συλλέγονται και να χρησιμοποιούνται στο πλαίσιο εξατομίκευσης του Ιστού. Τα στοιχεία αυτά ταξινομούνται σε τέσσερις κατηγορίες:

1. Περιεχομένου: δεδομένα που παρουσιάζονται στον χρήστη με κατάλληλη δομή (κείμενο, εικόνες, δομημένα δεδομένα που ανακτώνται από βάσεις δεδομένων κ.ο.κ.)

2. Δομής: δεδομένα που αντιπροσωπεύουν τον τρόπο οργάνωσης του περιεχομένου (HTML, XTML tags, hyperlinks κ.ο.κ.)

3. Χρήσης: δεδομένα που αντιπροσωπεύουν τη χρήση μιας ιστοσελίδας (IP επισκέπτη, ώρα και ημερομηνία πρόσβασης, πλήρης διαδρομή για αρχεία ή directories, διεύθυνση του συντάκτη κ.ο.κ.)

4. Προφίλ χρήστη: δεδομένα που παρέχουν πληροφορίες για τους χρήστες μιας ιστοσελίδας (δημογραφικές πληροφορίες, ενδιαφέροντα, προτιμήσεις κ.ο.κ.)

Η συνολική διαδικασία εξατομίκευσης του ιστού με βάση τη χρήση αποτελείται από πέντε υπομονάδες:

1. Δημιουργία προφίλ χρήστη: διαδικασία συλλογής πληροφοριών που αφορούν κάθε επισκέπτη, είτε άμεσα, είτε έμμεσα. Τα δεδομένα προφίλ χρήστη αξιοποιούνται για να προσαρμόσουν μια ιστοσελίδα στις συγκεκριμένες ανάγκες του επισκέπτη.

2. Ανάλυση αρχείων καταγραφής και εξόρυξη δεδομένων ιστού: διαδικασία όπου οι πληροφορίες που έχουν σημειωθεί υφίστανται επεξεργασία με σκοπό να εξάγουν στατιστικές πληροφορίες και να ανακαλύπτουν μοτίβα χρήσης, να ομαδοποιούν τους χρήστες σε ομάδες σύμφωνα με τη συμπεριφορά τους και να αποκαλύπτουν πιθανές συσχετίσεις μεταξύ ιστοσελίδων και ομάδων χρηστών. Η διαδικασία αυτή επικαλύπτεται με τη διαδικασία δημιουργίας προφίλ χρηστών.

3. Διαχείριση περιεχομένου: διαδικασία ταξινόμησης περιεχομένου σε σημασιολογικές κατηγορίες προκειμένου να γίνει ευκολότερη η ανάκτηση και η παρουσίαση πληροφοριών για τους χρήστες. Είναι σημαντική διαδικασία ιδιαίτερα για ιστοσελίδες των οποίων το περιεχόμενο αυξάνεται σε καθημερινή βάση.

4. Δημοσίευση ιστοσελίδας: μηχανισμός δημοσίευσης για την παρουσίαση του περιεχομένου που είναι τοπικά αποθηκευμένο σε έναν διακομιστή ιστού ή/και ορισμένων πληροφοριών που ανακτώνται από άλλους πόρους ιστού με ομοιόμορφο τρόπο στον τελικό χρήστη.

5. Απόκτηση και αναζήτηση πληροφοριών: τεχνική κατάταξης της αναζήτησης και της καταλληλότητας τόσο στη διαδικασία απόκτησης των σχετικών πληροφοριών όσο και στη δημοσίευση των κατάλληλων δεδομένων σε κάθε ομάδα χρηστών.

1.2 Εξατομίκευση Ιστού

Η εξατομίκευση ιστού μπορεί να οριστεί ως η διαδικασία προσαρμογής του περιεχομένου και της δομής μιας ιστοσελίδας που ανταποκρίνεται στις ιδιαίτερες ανάγκες κάθε χρήστη εκμεταλλευόμενη την συμπεριφορά περιήγησής του. Τα βήματα αυτής περιλαμβάνουν: α) τη συλλογή των δεδομένων ιστού, β) την κατηγοριοποίηση αυτών, γ) την ανάλυση αυτών και δ) τον προσδιορισμό των ενεργειών που πρέπει να εκτελεστούν. Η ιστοσελίδα εξατομικεύεται με την επισήμανση των υπαρχόντων hyperlinks, τη δυναμική εισαγωγή νέων hyperlinks που φαίνεται να ενδιαφέρουν τον τρέχοντα χρήστη, ή ακόμα και τη δημιουργία νέων index pages.

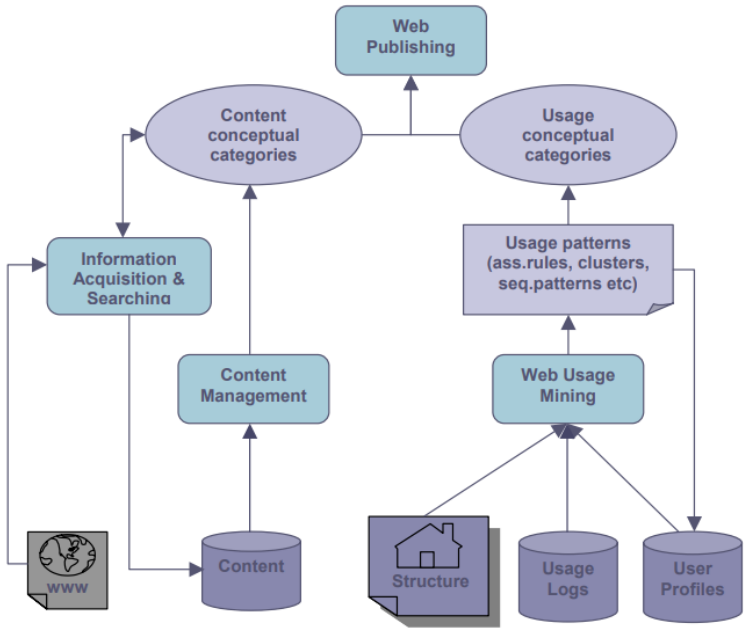
Τρόποι ανάλυσης δεδομένων:

1. Συστήματα φιλτραρίσματος βάσει περιεχομένου: βασίζονται αποκλειστικά στις προτιμήσεις των μεμονωμένων χρηστών.

2. Συνεργατικά συστήματα φιλτραρίσματος: καλούν τους χρήστες να αξιολογούν αντικείμενα ή να δηλώνουν προτιμήσεις και ενδιαφέροντα.

3. Συστήματα φιλτραρίσματος βάσει κανόνων: καλούν τους χρήστες να απαντήσουν σε ένα σύνολο ερωτήσεων.

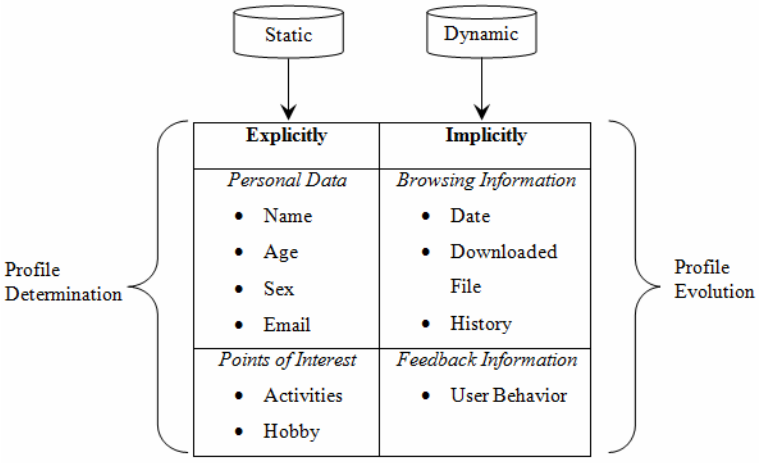
4. Συστήματα εξόρυξης δεδομένων από τη χρήση του ιστού: βασίζονται στην εφαρμογή στατιστικών μεθόδων και μεθόδων εξόρυξης δεδομένων στα δεδομένα του ιστολογίου, με αποτέλεσμα να διαμορφώνεται ένα σύνολο χρήσιμων μοτίβων που υποδεικνύουν τη συμπεριφορά των χρηστών κατά τη περιήγηση.



Εικόνα 1.1: Λειτουργικές μονάδες ενός συστήματος εξατομίκευσης του ιστού.

1.3 Δημιουργία προφίλ χρήστη

Το προφίλ ενός χρήστη μπορεί να είναι στατικό, όπου η πληροφορία του αλλάζει σπάνια ή σχεδόν ποτέ ή δυναμικό με τη πληροφορία αυτού να αλλάζει συχνά. Οι πληροφορίες αυτές λαμβάνονται είτε άμεσα, με χρήση φορμών καταχώρησης και ερωτηματολογίων (στατικό), είτε έμμεσα, με τη καταγραφή της συμπεριφορά περιήγησης ή των προτιμήσεων κάθε χρήστη ή με την ομαδοποίηση χρήστη.



Εικόνα 1.2: Στατική και δυναμική δομή του προφίλ. [8]

1.4 Συλλογή δεδομένων

Cookies: δεδομένα που στέλνει ένας διακομιστής ιστού σε έναν πελάτη ιστού, που αποθηκεύονται τοπικά από τον πελάτη και στέλνονται πίσω στον διακομιστή σε επόμενες αιτήσεις. Με άλλα λόγια ένα cookie είναι μια κεφαλίδα HTTP που

αποτελείται από μία συμβολοσειρά που εισάγεται στη μνήμη ενός προγράμματος περιήγησης. Χρησιμοποιείται για τη μοναδική αναγνώριση ενός χρήστη, κατά τη διάρκεια των αλληλεπιδράσεών του με τον ιστό μέσα σε μία τοποθεσία και περιέχει παραμέτρους δεδομένων που επιτρέπουν στον απομακρυσμένο HTML να διατηρεί ταυτότητας του χρήστη και των ενεργειών του. Γενικά, οι πληροφορίες του επισκέπτη αποθηκεύονται μαζί με πληροφορίες κωδικού πρόσβασης.

Identd: πρωτόκολλο αναγνώρισης που παρέχει προσδιορισμό ταυτότητας ενός χρήστη μιας συγκεκριμένης TCP διεύθυνσης. Με δεδομένο ένα ζεύγος αριθμών θυρών TCP, επιστρέφει μια συμβολοσειρά χαρακτήρων που προσδιορίζει των ιδιοκτήτη αυτής της σύνδεσης στο σύστημα του διακομιστή ιστού.

Τέλος, οι χρήστες μπορούν να ταυτοποιηθούν βάσει της παραδοχής ότι σε κάθε IP αντιστοιχεί και ένας χρήστης. Σε συγκεκριμένες περιπτώσεις, οι διευθύνσεις χωρίζονται σε ονόματα τομέων που καταχωρίζονται σε ένα άτομο ή μία εταιρεία.

Η ανάκτηση δεδομένων χρήστη μέσω cookies, δεν είναι δυνατή στην περίπτωση που ο χρήστης απενεργοποιήσει την χρήση τους. Το γεγονός ότι οι πληροφορίες αυτές αποθηκεύονται τοπικά, δίνει την δυνατότητα στον χρήστη να τις σβήσει, με αποτέλεσμα, όταν επανεπισκεφτεί την τοποθεσία να θεωρηθεί νέος επισκέπτης. Αν δεν πιστοποιηθεί κωδικός πρόσβασης υπάρχει περίπτωση χρήσης πάνω από ένα ατόμων. Παρόμοια μειονεκτήματα υπάρχουν και στην μέθοδο identd και ip address resolving. Επιπλέον, οι φόρμες συμπλήρωσης και τα ερωτηματολόγια μπορούν να έχουν εσφαλμένες πληροφορίες που μπορούν να οδηγήσουν σε λανθασμένα προφίλ χρηστών. Τέλος, δημιουργούνται θέματα παραβίασης απορρήτου.

1.5 Ανάλυση αρχείων καταγραφής και εξόρυξη στοιχείων χρήσης ιστού

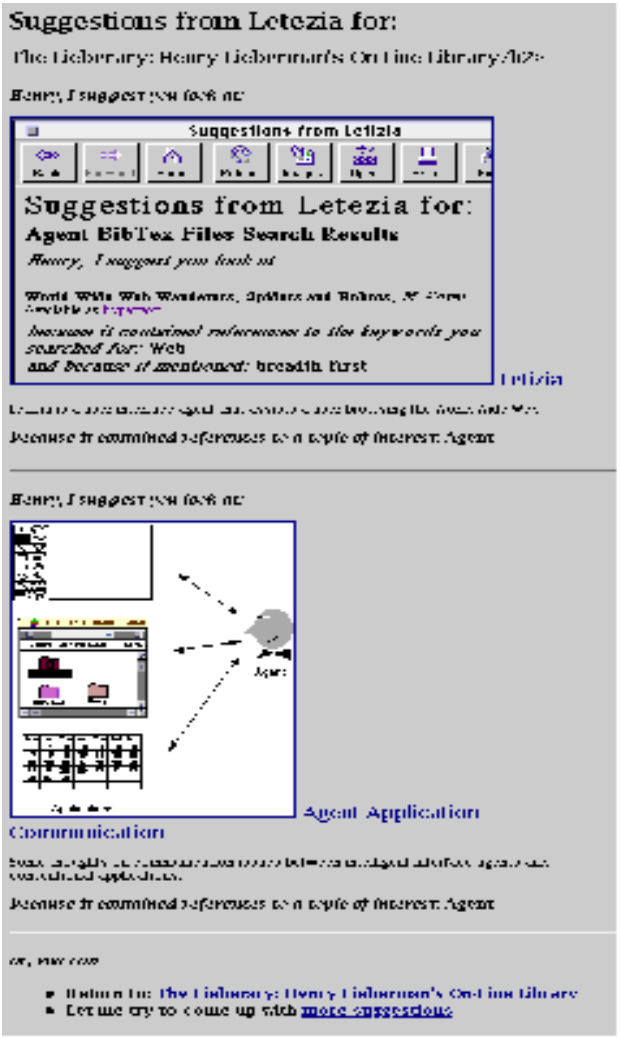
Η εξόρυξη δεδομένων χρήσης του ιστού είναι διαδικασία εφαρμογής στατιστικών μεθόδων και εξόρυξης δεδομένων σε δεδομένα ιστολογιών, προκειμένου να εξαχθούν χρήσιμα μοτίβα που αφορούν τη συμπεριφορά των χρηστών κατά τη περιήγηση, τις συστοιχίες χρηστών και σελίδων, καθώς και πιθανές συσχετίσεις μεταξύ ιστοσελίδων και ομάδων χρηστών.

Οι κανόνες και τα μοτίβα που εξάγονται μπορούν, στη συνέχεια, να χρησιμοποιηθούν για τη βελτίωση της απόδοσης του συστήματος ή για τροποποιήσεις της ιστοσελίδας. Οι πληροφορίες που περιλαμβάνονται στα αρχεία καταγραφής μπορούν να ενσωματωθούν στα δεδομένα πελατών που συλλέγονται από συστήματα CRM και ERP, προκειμένου να συγκεντρωθούν επιχειρηματικές πληροφορίες. Τα συστήματα ERP παρέχουν απρόσκοπτη ενοποίηση των διαδικασιών σε λειτουργικούς τομείς με βελτιωμένη, ροή εργασιών, τυποποίηση διαφόρων επιχειρηματικών πρακτικών, βελτιωμένη διαχείριση παραγγελιών, ακριβή λογιστική των αποθεμάτων, και καλύτερη διαχείριση εφοδιαστικής αλυσίδας [9]. Τα συστήματα ERP περιλαμβάνουν όλες τις πτυχές της επικοινωνίας και της ενασχόλησης ενός οργανισμού με τον πελάτη του, ανεξάρτητα από το αν συνδέεται με προϊόντα ή υπηρεσίες. [10]

Πρέπει να ληφθούν υπόψη πολλά θέματα, όπως οι αποφάσεις που πρέπει να λαμβάνονται κατά το φιλτράρισμα και την επεξεργασία των δεδομένων, η ταυτότητα χρήστη και περιόδου λειτουργίας και η ταυτότητα του pageview.

Επίσης σημαντική είναι και η επιλογή μεθόδων εξόρυξης.

1.6 Ερευνητικές Πρωτοβουλίες



Lieberman (1995): Letizia, μία από τις πρώτες προσπάθειες αξιοποίησης πληροφοριών περιήγησης ενός επισκέπτη, παρακολουθεί τη συμπεριφορά του χρήστη κατά τη περιήγησή του και αναζητά best-first με τη χρήση ευρετικών μεθόδων πιθανώς ενδιαφέρουσες σελίδες για συστάσεις.

Εικόνα 1.3: Παράδειγμα χρήσης της Letizia.

Yan (1996): προσέγγιση για αυτόματη ταξινόμηση επισκεπτών σύμφωνα με τα μοτίβα πρόσβασής τους. Το μοντέλο αποτελείται από δύο ενότητες: μια λειτουργική μονάδα εκτός σύνδεσης που εκτελεί ανάλυση συστάδων στα αρχεία καταγραφής και μία λειτουργική μονάδα σε σύνδεση με σκοπό τη δημιουργία δυναμικής σύνδεσης. Κάθε χρήστης αντιστοιχίζεται σε μία συστάδα με βάση τα τρέχοντα μοτίβα διέλευσης (offline υλοποίηση).

Joachims (1999): WebWatcher, ξεναγός που παρέχει συμβουλές περιήγησης στον χρήστη μέσω μιας δεδομένης συλλογής ιστού, με βάση τις γνώσεις του για τα ενδιαφέροντα του χρήστη, τη θέση και τη σχετικότητα των διαφόρων αντικειμένων και τον τρόπο αλληλεπίδρασής του στο παρελθόν. Η στρατηγική του για την παροχή συμβουλών μαθαίνεται από ανατροφοδότηση από προηγούμενες περιηγήσεις. Ένα παρόμοιο σύστημα είναι το Personal WebWatcher (Mladenic 1999), το οποίο δομείται στην εξειδίκευση για ένα συγκεκριμένο χρήστη. Καταγράφει αποκλειστικά τις διευθύνσεις των σελίδων που ζητούνται από τον χρήστη και επισημαίνει τα

ενδιαφέροντα hyperlinks χωρίς να εμπλέκει τον χρήστη στη διαδικασία εκμάθησης, ζητώντας λέξεις.

Σπηλιοπούλου (1998-2000): MINT, γλώσσα εξόρυξης για την εφαρμογή του WUM, ένα σύστημα εξόρυξης ακολουθίας για την προδιαγραφή, ανακάλυψη και απεικόνιση των ενδιαφερόντων μοτίβων πλοήγησης. Γίνεται προεπεξεργασία του αρχείου καταγραφής και αποθηκεύεται μια συνολική υλική άποψη του ιστολογίου. Στη φάση προετοιμασίας των δεδομένων, εκτός από το φιλτράρισμα και την ολοκλήρωση των δεδομένων του αρχείου καταγραφής, οι περίοδοι λειτουργίας χρήστη προσδιορίζονται με τη χρήση μηχανισμών χρονικών ορίων. Η διαδρομή που ακολουθεί κάθε χρήστης ονομάζεται «μονοπάτι». Επειδή πολλοί χρήστες έχουν πρόσβαση στις ίδιες σελίδες με την ίδια σειρά, ένα «δέντρο συγκεντρωτικών αποτελεσμάτων» είναι κατασκευασμένες με συγχώνευση μονοπατιών με το ίδιο πρόθεμα. Αυτό το δέντρο ονομάζεται «συγκεντρωτικό ημερολόγιο» και τα μοτίβα πλοήγησης που παρουσιάζουν ενδιαφέρον μπορούν να εξαχθούν με MINT.

Mobasher (1999,2000): WebPersonaliser παρέχει ένα πλαίσιο για την εξόρυξη των αρχείων καταγραφής ιστού για να ανακαλύψει τη γνώση για την παροχή των συστάσεων στους τρέχοντες χρήστες με βάση τους περιήγηση ομοιότητες με τους προηγούμενους χρήστες. Βασίζεται αποκλειστικά σε ανώνυμα δεδομένα χρήσης που παρέχονται από κορμούς και τη δομή υπερκειμένου μιας τοποθεσίας. Μετά τη συλλογή και την προεπεξεργασία δεδομένων (μετατροπή των πληροφοριών χρήσης, περιεχομένου και δομής που περιέχονται στις διάφορες πηγές δεδομένων σε διάφορες αφαιρέσεις δεδομένων), εφαρμόζονται τεχνικές εξόρυξης δεδομένων, όπως κανόνες συσχέτισης, διαδοχική ανακάλυψη μοτίβων, δημιουργία συμπλεγμάτων και ταξινόμηση, προκειμένου να ανακαλυφθούν ενδιαφέροντα μοτίβα χρήσης. Στη συνέχεια, τα αποτελέσματα χρησιμοποιούνται για τη δημιουργία συγκεντρωτικών προφίλ χρήσης, προκειμένου να δημιουργηθούν κανόνες απόφασης. Η μηχανή προτάσεων αντιστοιχίζει τη δραστηριότητα κάθε χρήστη με αυτά τα προφίλ και του παρέχει μια λίστα με τις προτεινόμενες συνδέσεις υπερκειμένου.

Project Name	Data Source	User Profiling	Web Usage Mining	Content Management	Publishing Mechanism
Letizia [Lieberman 1995]	Client		*	(*)	
WebWatcher [Joachims et al. 1997]	Proxy	*	*	(*)	
Analog [Yan et al. 1996]	Server		*		* (Suggested)
SpeedTracer [Wu et al. 1998]	Server		*		
WebLogMiner [Zaiane et al. 1998]	Server		*		
Borges and Levene [1999]	Server		*		
Shahabi et al. [1997]	Client	*	*		
Joshi et al. [2000; Krishnapuram et al. 2001; Nasraoui et al. 2000]	Server	*	*		
WebSIFT [Cooley et al. 1999b,a; Srivastava et al. 2000]	Server		*	(*)	
WebTool [Masseglia et al. 1999a,b, 2000]	Server		*		* (Suggested)
Buchner et al. [Buchner and Mulvenna 1998; Buchner et al. 1999]	Server	*	*	*	
WUM [Spiliopoulou and Faulstich 1998; Spiliopoulou et al. 1999; Spiliopoulou 2000]	Server		*		
STRATDYN [Berendt 2000, 2001]	Server		*	*	
Coenen et al. [2000]	Server		*		* (Suggested)
Adaptive Web Sites [Perkowitz and Etzioni 1999, 2000]	Server		*	*	*
Cingil et al. [2000]	Client		*	*	* (Suggested)
WebPersonalizer [Mobasher et al. 1999, 2000a]	Server	*	*		*
Mobasher et al. [2000b,c]	Server	*	*	*	*

Εικόνα 1.4: Ερευνητικές Πρωτοβουλίες όπως αναλύονται στο άρθρο [4].

2. Μοντέλο πρόβλεψης ιστοσελίδων που βασίζεται στην αναπαράσταση δέντρου click-stream της συμπεριφοράς του χρήστη (A Web Page Prediction Model Based on Click-Stream Tree Representation of User Behavior)

2.1 Πρόβλεψη συμπεριφοράς χρήστη

Η εξόρυξη δεδομένων ιστού ορίζεται ως η χρήση τεχνικών εξόρυξης δεδομένων για τον αυτόματο εντοπισμό και εξαγωγή πληροφοριών από τα έγγραφα και τις υπηρεσίες του ιστού. Με την ανάπτυξη του παγκόσμιου ιστού, η μελέτη πρόβλεψης ιστοσελίδων για τον χρήστη είναι πιο σημαντική. Η διαδικασία αυτή χωρίζεται σε τρία βήματα: 1. Καθαρισμός δεδομένων και προετοιμασία για την εξόρυξη μοτίβων χρήσης, 2. Εξαγωγή μοτίβων χρήσης και 3. Δημιουργία μοντέλου πρόβλεψης βάσει των εν λόγω μοτίβων. Οι κύριες τεχνικές που χρησιμοποιούνται παραδοσιακά στη δημιουργία μοντέλων για τα μοτίβα χρήσης μια τοποθεσίας ιστού είναι το

συνεργατικό φιλτράρισμα (collaborative filtering, CF)-λαμβάνει υπόψη του όχι μόνο τον επικείμενο χρήστη αλλά και άλλους- [11], η δημιουργία σελίδων συμπλέγματος ή περιόδων λειτουργίας χρήστη, η δημιουργία κανόνα συσχέτισης, η δημιουργία διαδοχικών μοτίβων και μοντέλων Markov. Το στάδιο της πρόβλεψης είναι η επεξεργασία του μοντέλου σε πραγματικό χρόνο, η οποία εξετάζει την ενεργή συνεδρία του χρήστη και κάνει συστάσεις με βάση τα εντοπισμένα μοτίβα.

Ωστόσο, τα μοτίβα που αναφέρθηκαν δεν είναι επαρκή για την επακριβή περιγραφή της συμπεριφοράς του χρήστη στον ιστό, καθώς δεν λαμβάνουν υπόψη τον χρόνο που αυτός καταναλώνει στην εκάστοτε ιστοσελίδα. Ο αφιερωμένος χρόνος είναι ενδεικτικός του ενδιαφέροντος του χρήστη και επιπλέον, ο ίδιος ο χρήστης μπορεί να έχει διαφορετικές ανάγκες και επιτυχίες βάσει χρονοδιαγράμματος. Το μοντέλο που προτείνεται στο άρθρο [4], λαμβάνει υπόψη τα μοτίβα χρήσης καθώς και τον χρόνο.

2.2 Προτεινόμενο Μοντέλο

2.2.1. Προετοιμασία δεδομένων και καθαρισμός

Στο πείραμα της προκείμενης μελέτης, χρησιμοποιήθηκαν δύο σύνολα δεδομένων διακομιστή: το πρώτο είναι μέρος NASA Kennedy Space Center server (Ιούλιος και Αύγουστος του 1995) και το δεύτερο του ClarkNet Web server, έναν πλήρη πάροχο πρόσβασης στο διαδίκτυο για την περιοχή Metro Baltimore-Washington DC (Αύγουστος και Σεπτέμβριος του 1995).

Οι ώρες επίσκεψης σελίδας, οι οποίες εξάγονται κατά τη διαδικασία καθαρισμού, κανονικοποιούνται κατά τις ώρες επίσκεψης των σελίδων στην ίδια συνεδρία, έτσι ώστε η ελάχιστη τιμή κανονικοποιημένου χρόνου είναι 1. Για την αξιολόγηση του αποτελέσματος των τιμών κανονικοποίησης, δοκιμάζουμε 4 διαφορετικές μέγιστες τιμές: 2, 3, 5 και 10. Αν μια σελίδα δεν βρίσκεται στην περίοδο λειτουργίας χρήστη, τότε η τιμή του αντίστοιχου κανονικοποιημένου χρόνου ορίζεται σε 0. Αυτή η κανονικοποίηση αποτυπώνει τη σχετική σπουδαιότητα μιας σελίδας σε έναν χρήστη σε μια περίοδο λειτουργίας.

2.2.2. Μέτρηση ομοιότητας περιόδου λειτουργίας

Προτείνεται ένα μέτρο ομοιότητας συνεδρίας που βασίζεται στη μέθοδο ευθυγράμμισης ακολουθίας FastLSA -επεκτείνει τα όρια ανάλυσης τοπικής ομοιότητας [12]. Επειδή οι περίοδοι χρήσης είναι διατεταγμένες σε URL, μπορούμε να τις αποκαλούμε ακολουθίες ιστοσελίδων. Το πρόβλημα της εύρεσης της βέλτιστης στοίχισης ακολουθίας λύνεται με τη χρήση δυναμικού προγραμματισμού. Ο αλγόριθμος χρησιμοποιεί μια μήτρα στην οποία η μία ακολουθία τοποθετείται κατά μήκος της κορυφής της μήτρας και η άλλη κατά μήκος της αριστερής πλευράς της μήτρας. Στο τέλος κάθε ακολουθίας, προστίθεται ένα κενό το οποίο υποδεικνύει το σημείο εκκίνησης για τον υπολογισμό της βαθμολογίας ομοιότητας. Η δισδιάστατη βαθμολογία υπολογίζονται για ένα ζεύγος σελίδων που ταιριάζουν.

	P_4	P_1	P_2	P_5	P_3	P_6	-
P_1	2	**3**	0	-3	-4	-4	-4
P_2	-1	0	1	-2	-3	-3	-3
P_4	-1	-2	-1	-1	-2	-2	-2
P_5	-3	-2	-1	0	-2	-1	-1
-	-6	-5	-4	-3	-2	-1	0

Εικόνα 2.1: Η μήτρα βαθμολόγησης για ακολουθίες δύο διαστάσεων

Κάθε ασυμφωνία ή κενό που εισάγεται στις ακολουθίες βαθμολογείται με πέναλτι -1. Η ομοιότητα υπολογίζεται, έπειτα, έτσι ώστε μόνο το ίδιο ταίριασμα ομοιότητας να έχει τιμή ομοιότητας 1. Το μέτρο αυτό αποτελείται από τη συνιστώσα βαθμολογίας ευθυγράμμισης -πόσο παρόμοιες είναι δύο συναρτήσεις στην περιοχή επικάλυψής τους- και από τη συνιστώσα τοπικής ομοιότητας.

2.2.3 Συσταδοποίηση ανά ζεύγη

Παίζει έναν εξαιρετικά σημαντικό ρόλο σε ολόκληρη τη διαδικασία, καθώς η κατηγοριοποίηση των δεδομένων είναι ένα από τα στοιχειώδη βήματα στην ανακάλυψη της γνώσης. Πρόκειται για μια εργασία μάθησης χωρίς επίβλεψη που χρησιμοποιείται για διερευνητική ανάλυση δεδομένων ώστε να βρεθούν ορισμένα μη αποκαλυπτικά μοτίβα που δεν δύνανται να ταξινομηθούν με σαφήνεια [13]. Κατασκευάζεται γράφημα με κορυφές τις περιόδους χρήσης. Υπάρχει ακμή μεταξύ δύο κορυφών, αν η τιμή ομοιότητας είναι μεγαλύτερη από 0 και αυτή η ακμή είναι σταθμισμένη με αυτήν της ομοιότητας, τότε οι κορυφές είναι περίοδοι χρήσης.

2.2.4 Αναπαράσταση συστάδων

Οι συστάδες που παράγονται από τον αλγόριθμο διαμέρισης γραφήματος περιέχουν περιόδους χρήσης. Κάθε περίοδος χρήσης σε μία συστάδα είναι μία ακολουθία ιστοσελίδων που επισκέπτονται ένας χρήστης και ο μη υπολογισμένος χρόνος που δαπανάται σε αυτές τις σελίδες με ένα μοναδικό αριθμό περιόδου χρήσης. Δημιουργείται click-stream δέντρο για κάθε συστάδα. Κάθε δέντρο έχει έναν κόμβο ρίζας με ετικέτα null. Κάθε κόμβος, εκτός του ριζικού, αποτελείται από δεδομένα– αριθμό σελίδας και κανονονικοποιημένες πληροφορίες ώρας της σελίδας–, πλήθος – πλήθος περιόδων χρήσης που αντιπροσωπεύονται από το τμήμα διαδρομής που φτάνει σε κάθε κόμβο– και επόμενο κόμβο. Κάθε δέντρο διαθέτει πίνακα δεδομένων που αποτελείται από το πεδίο δεδομένων και τον πρώτο κόμβο. Το δέντρο κατασκευάζεται με τον αλγόριθμο της εικόνας 2.

1: Create a *root* node of a click-stream tree, and label it as "null"
2: $index \leftarrow 0$
3: **while** $index \leq$ *number of Sessions in the cluster* **do**
4: $Active\_Session \leftarrow t_{index}$
5: $m \leftarrow 0$
6: $Current\_Node \leftarrow$ *root* node of the click-stream tree
7: **while** $m \leq Active\_Session$ length **do**
8: $Active\_Data \leftarrow \{p^m_{t_{index}}\}\text{-}\{T^m_{pt_{index}}\}$
9: **if** there is a *Child* of $Current\_Node$ with the same data field **then**
10: $Child.count++$
11: $Current\_Node \leftarrow Child$
12: **else**
13: create a child node of the $Current\_Node$
14: $Child.data = Active\_Data$
15: $Child.count = 1$
16: $Current\_Node \leftarrow Child$
17: **end if**
18: $m++$
19: **end while**
20: $index++$
21: **end while**

Εικόνα 2.2: Αλγόριθμος κατασκευής δέντρου click-stream.

Τα παραγόμενα δέντρα χρησιμοποιούνται για προτάσεις.

2.2.5 Μηχανή προτάσεων

Η μηχανή προτάσεων είναι το στοιχείο πραγματικού χρόνου που επιλέγει την βέλτιστη διαδρομή για την πρόβλεψη της επόμενης αίτησης της περιόδου χρήσης του ενεργού χρήστη.

```
1: t_a ← Active User Session
2: if t_a.length ≤ 2 then
3:    Clusters = All Clusters
4: else
5:    Clusters = Top − N Clusters
6: end if
7: for i = 0 to NumberOfClusters do
8:    cl = Clusters[i]
9:    Sim[cl] = 0
10:   d_a ← {p^m_{t_a}}_{T_{p^m_{t_a}}}
11:   Node ← data_table[cl](d_a).first_node
12:   path = null
13:   while Node ≠ null do
14:      path = {path} + {Node.data}
15:      Parent_Node ← Node.Parent
16:      while Parent_Node ≠ null do
17:         path = {path} + {Parent_Node.Data}
18:         Parent_Node ← Parent_Node.Parent
19:      end while
20:      Sim(path) = sim(t_a, path)∗ Node.count/S[cl]
21:      if Sim(path) > Sim[cl] then
22:         Sim[cl] ← Sim(path)
23:         BestPath[cl] ← path
24:      end if
25:      path = null
26:      Node ← Node.next_node
27:   end while
28: end for
29: if t_a.length = 2 then
30:   Top−N Clusters ← N Clusters with highest Sim[cl]
      values
31: end if
```

Εικόνα 2.3: Αλγόριθμος εύρεσης καλύτερης διαδρομής.

2.3 Αποτελέσματα πειράματος

Με τη χρήση του single click-stream δέντρου αποκτήθηκαν τα βέλτιστα αποτελέσματα, δηλαδή το άνω όριο της ακρίβειας πρόβλεψης. Με τον τρόπο αυτό, δεν παρουσιάστηκαν τα προβλήματα που προκύπτουν με τη χρήση αλγορίθμων συσταδοποίησης.

Γενικά, η ακρίβεια πρόβλεψης και ο χρόνος που χρειάζεται για τις προτάσεις είναι σχετικώς αντιστρόφως ανάλογα. Για αυτόν τον λόγο, καθορίστηκαν μόνο οι N καλύτερες συστάδες έπειτα από δύο αιτήσεις. Τα πειράματα επαναλήφθηκαν αρκετές φορές και με διαφορετικές τιμές παραμέτρων, ώστε να επιτευχθεί αξιοπιστία.

Οι ερευνητές κατέληξαν στην μη αναγκαιότητα χρήσης της χρονικής πληροφορίας, καθώς με αυτόν τον τρόπο, προέκυψαν καλύτερα αποτελέσματα στις μετρικές αφού πια στο δέντρο υπάρχει μόνο η πληροφορία σχετικά με τον αριθμό της σελίδας.

Οι σημαντικότερες μετρικές που λήφθηκαν υπόψιν αποτελούν το hit-ratio και το click-soon-ratio. Ένα hit δηλώνεται αν οποιαδήποτε από τις τρεις προτεινόμενες σελίδες είναι η επόμενη αίτηση του χρήστη. Ο λόγος αυτός είναι ο αριθμός των επισκέψεων διαιρεμένος με τον συνολικό αριθμό των προτάσεων που γίνονται από το σύστημα. Ένα click-soon δηλώνεται αν οποιαδήποτε από τις τρεις προτεινόμενες σελίδες ζητηθεί από τον χρήστη κατά την ενεργή περίοδο χρήσης. Ο λόγος αυτός είναι ο αριθμός των click-soon διαιρεμένος με το συνολικό αριθμό προτάσεων που γίνονται από το σύστημα.

No.Of Clusters	Top-N					
	1		2		3	
	H-R	CS-R	H-R	CS-R	H-R	CS-R
5	56.19	91.17	57.23	92.82	57.95	93.89
10	53.92	88.31	55.07	89.9	56.01	91.27
15	52.3	86.36	53.77	88.21	54.68	89.36
20	48.96	80.69	50.58	83.01	52.10	84.20
25	48.67	80.15	50.02	82.45	50.42	82.98
30	48.33	79.50	49.37	81.17	50.58	82.74

Εικόνα 2.4: Εκατοστιαία αποτελέσματα στο σύνολο δεδομένων NASA, όταν η χρονική πληροφορία αγνοείται.

No.Of Clusters	Top-N					
	1		2		3	
	H-R	CS-R	H-R	CS-R	H-R	CS-R
5	56.19	91.17	57.23	92.82	57.95	93.89
10	53.92	88.31	55.07	89.9	56.01	91.27
15	52.3	86.36	53.77	88.21	54.68	89.36
20	48.96	80.69	50.58	83.01	52.10	84.20
25	48.67	80.15	50.02	82.45	50.42	82.98
30	48.33	79.50	49.37	81.17	50.58	82.74

Εικόνα 2.5: Εκατοστιαία αποτελέσματα στο σύνολο δεδομένων ClarkNet, όταν η χρονική πληροφορία αγνοείται.

3. Έγκαιρος εντοπισμός εξόδων χρήστη από δεδομένα click-stream : Ένα Markov Μοντέλο Διεργασίας Με Διαμόρφωση Μαρκοειδών Σημείων (Early Detection of User Exits from Clickstream Data: A Markov Modulated Marked Point Process Model)

Οι περισσότεροι χρήστες αφήνουν ιστοσελίδες ηλεκτρονικού εμπορίου χωρίς να αγοράσουν κάποιο προϊόν . Είναι σημαντικό για τους ιδιοκτήτες ιστοτόπων να ανιχνεύσουν τους χρήστες που κινδυνεύουν να εξέλθουν και να παρέμβουν νωρίς. Για το λόγο αυτό αναπτύχθηκε ένα νέο Markov μοντέλο εντοπισμού χρηστών που ενδέχεται να βρεθούν χωρίς κάποια αγορά και ένα πλαίσιο αξιολόγησης κινδύνου. Τα υπολογιστικά πειράματα βασίζονται σε πραγματικά δεδομένα ροής “κλικ”. Με

βάση αυτό, διαπιστώνουμε ότι οι κορυφαίοι αλγόριθμοι ξεπερνιούνται από το προτεινόμενο μοντέλο M3PP όσον αφορά τόσο AUROC (+6.24 ποσοστιαίες μονάδες) όσο και τον αποκαλούμενο χρόνο έγκαιρης προειδοποίησης (+12.93 %). Αντίστοιχα, το M3PP όσο αναφορά για έγκαιρες ανιχνεύσεις των εξόδων των χρηστών παρέχει επαρκή χρόνο στους ιδιοκτήτες ιστοτόπων ηλεκτρονικού εμπορίου για την ενεργοποίηση δυναμικών ηλεκτρονικών παρεμβάσεων. [14], [15], [16]

1. Μοντέλο M3PP: Πρόκειται για ένα νέο μοντέλο M3PP1 που μοντελοποιεί τόσο την ακολουθία των μεμονωμένων σελίδων που επισκέπτεται και τον χρόνο που δαπανάται στις σελίδες. Τυπικά, αυτό το έργο απαιτεί ένα προσαρμοσμένο μοντέλο: αντί να υποθέτουμε ότι το clickstream είναι μια διαδικασία διακριτού χρόνου με βήματα μονάδας, η ρύθμισή απαιτεί μια διαδικασία συνεχούς χρόνου. Ο κίνδυνος εξόδου χωρίς καμία αγορά, επιτυγχάνει κορυφαία απόδοση.

2. Πλαίσιο αξιολόγησης κινδύνου: Πρόκειται για ένα νέο πλαίσιο αξιολόγησης κινδύνου για τον εντοπισμό χρηστών που κινδυνεύουν να εξέλθουν χωρίς αγορά. Παρέχεται, αρχικά, μια βαθμολογία κινδύνου που υπολογίζει την πιθανότητα ενός χρήστη να εξέλθει τελικά χωρίς αγορά σε πραγματικό χρόνο. Βάσει του πλαισίου η απόδοση του μοντέλου αξιολογείται με τυποποιημένες μετρήσεις (ROC) και μετρικό που ονομάζεται χρόνος έγκαιρης προειδοποίησης, η οποία επικεντρώνεται στην έγκαιρη ανίχνευση των χρηστών σε κίνδυνο. Ο έγκαιρος εντοπισμός είναι κρίσιμος για την παροχή επαρκής χρόνος για την ενεργοποίηση επιτυχημένων παρεμβάσεων.

Το 2018, πάνω από το 97% των χρηστών του παγκόσμιου ηλεκτρονικού εμπορίου βγήκαν από τον ιστότοπο χωρίς αγορά. Εξαιτίας αυτού, υπάρχουν σημαντικές δυνατότητες για τους διαδικτυακούς λιανοπωλητές να αυξήσουν τα έσοδά τους. Η ανίχνευση των χρηστών από τα δεδομένα του clickstream είναι μια δύσκολη επιχείρηση. Ο στόχος στην πράξη δεν είναι μόνο ανίχνευση ακριβώς πριν από τις εξόδους τους, αλλά η ανίχνευση των χρηστών που κινδυνεύουν να εξέλθουν σε πρώιμο στάδιο της συνεδρίασής τους. Τέτοιες παρεμβάσεις μπορούν να έχουν διάφορες μορφές, για παράδειγμα, παρουσίαση περιεχομένου μάρκετινγκ, δυναμική προσαρμογή σχεδιασμού ιστοσελίδας ή προσφορά προώθησης τιμών, με στόχο την καθοδήγηση των χρηστών που διαφορετικά θα είχαν εξέλθει χωρίς αγορά προς μετατροπή δηλαδή μετατροπή ενός χρήστη σε αγοραστή.

Έχουν προταθεί αρκετές προσεγγίσεις για τη μοντελοποίηση των δεδομένων clickstream με ιδιαίτερα δημοφιλή για το χειρισμό διαδοχικών δεδομένων : όπως η βαθιά μάθηση και το Markov μοντέλο.

- Μοντέλα Markov: Με βάση το μοντέλο Markov οι σελίδες τις οποίες έχει επισκεφτεί ένας χρήστης είναι προβλέψιμες. Ωστόσο, η οικογένεια των αλυσίδων Markov για διακριτή χρονική περίοδο υποθέτει ότι η διαδικασία εξελίσσεται σε βήματα ανά μονάδα χρόνου. Επομένως, ο χρόνος που δαπανάται στη σελίδα αγνοείται. Αν και οι αλυσίδες Markov υψηλότερου μήκους και οι αλυσίδες Markov μεταβλητού μήκους είναι συνήθως πιο ακριβείς για την πρόβλεψη των clickstreams των χρηστών, έρχονται με υπολογιστικό κόστος λόγω του εκθετικά μεγάλου χώρου τους. Επιπλέον, η αύξηση του αριθμού των καταστάσεων μπορεί ακόμη και να οδηγήσει σε χειρότερη απόδοση και μπορεί να περιορίσει σημαντικά τη χρησιμότητα τους για εφαρμογές που απαιτούν γρήγορες προβλέψεις, όπως η συναγωγή του κινδύνου ενός χρήστη να εξέλθει χωρίς αγορές σε πραγματικό χρόνο.

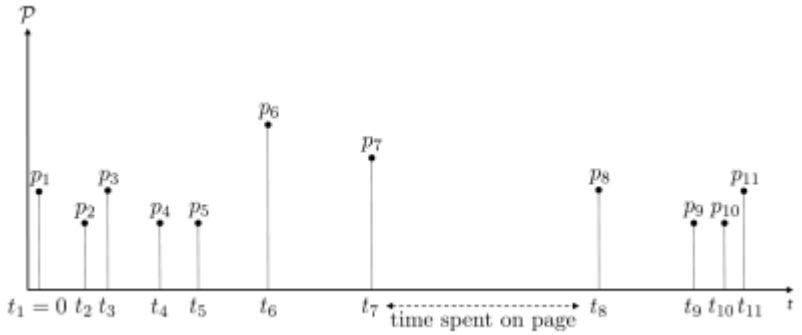
- Κρυμμένα μοντέλα Markov: Τα μοντέλα Hidden Markov (HMMs) περιγράφουν τις λανθάνουσες φάσεις αγορών του χρήστη με τη χρήση μιας αλυσίδας Markov διακριτού χρόνου με τις σελίδες που επισκέφθηκαν να είναι οι εκπομπές που εξαρτώνται από την τρέχουσα φάση αγορών. [17], [18]

- Διεργασίες σημαδεμένων σημείων: Εκτός από τα μοντέλα Markov, αναθεωρούμε τις διαδικασίες σημαδεμένων σημείων, οι οποίες έχουν γίνει δημοφιλείς λόγω της ικανότητάς τους να καταγράφουν τόσο τον χρόνο των γεγονότων όσο και τον τύπο του συμβάντος. Ωστόσο, δεν έχουν προσαρμοστεί ακόμα στα δεδομένα clickstream.

- Κενό στην έρευνα: Η οικογένεια των αλυσίδων Markov διακριτού χρόνου δεν μπορεί να αποτυπώσει τόσο την ακολουθία των σελίδων που επισκέπτεται όσο και τους χρόνους που δαπανώνται στη σελίδα, αφού θεωρεί ότι η διαδικασία εξελίσσεται με βήματα της μονάδας. Ως εκ τούτου, αυτό απαιτεί μια οικογένεια μοντέλων που μπορούν να αντιμετωπίσουν τις διαδικασίες που εξελίσσονται σε συνεχή χρόνο. Επιπλέον, οι προηγούμενες εργασίες δεν έχουν ασχοληθεί με τον έγκαιρο εντοπισμό του χρήστη που κινδυνεύει να βγει από το κατάστημα χωρίς να γίνει αγορά. [19], [20]

  Δομή ενός συνόλου δεδομένων clickstream: Ένα σύνολο δεδομένων clickstream αποτελείται από ένα σύνολο περιόδων εργασίας· κάθε περίοδος λειτουργίας είναι μια ακολουθία σελίδων που έχει επισκεφθεί ένας χρήστης μαζί με μια χρονική σήμανση κατά την επίσκεψή του. Επίσης, οι περίοδοι λειτουργίας μπορεί να έχουν διάφορα μήκη.

Υποδηλώνουμε ένα σύνολο δεδομένων clickstream ως D και τις περιόδους λειτουργίας ως : $\mathrm{D} = \{\mathrm{S_d}\}_{\mathrm{d=1}}^{\mathrm{D}}$ όπου s^d είναι η d-η επίσκεψη του χρήστη.
Ο συνολικός αριθμός σελίδων που παρατηρήθηκαν κατά τη διάρκεια αυτής της επίσκεψης δηλώνεται από Md και η διάρκεια της συμβολίζεται με T^d.



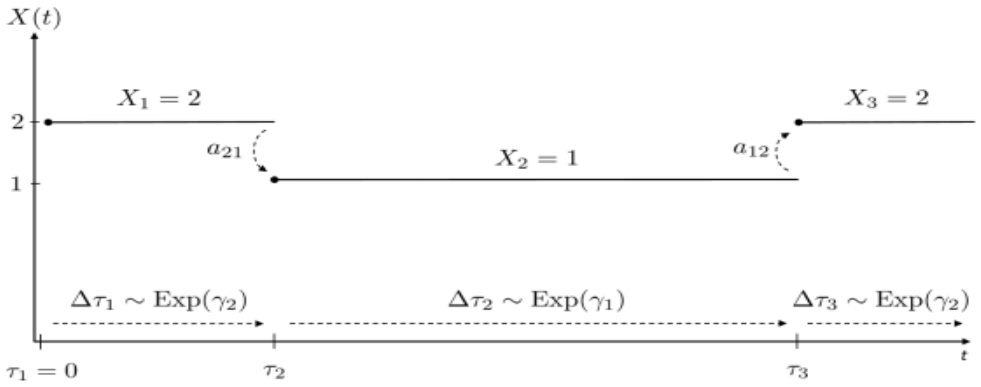
Εικόνα 3.1: Παράδειγμα σελίδων που επισκέφτηκε ο χρήστης συμπεριλαμβανομένου του χρόνου που δαπάνησε σε κάθε σελίδα

Διαδικασία M3PP:

- Υλοποίηση μιας διαδικασίας Markov jump process :

$$X(t) = \sum_{n=1}^{k} x_n \ 1\{\tau_n \leq t < \tau_n + 1\}$$

Χρησιμοποιούμε μια διαδικασία Markov jump X(t) για να μοντελοποιήσουμε τις λανθάνουσες φάσεις αγορών ενός χρήστη. Είναι μια συνεχής χρονική επέκταση μιας αλυσίδας Markov διακριτού χρόνου. Η υλοποίησή της είναι μια τμηματικά σταθερή λειτουργία μετάβασης μεταξύ των φάσεων N.



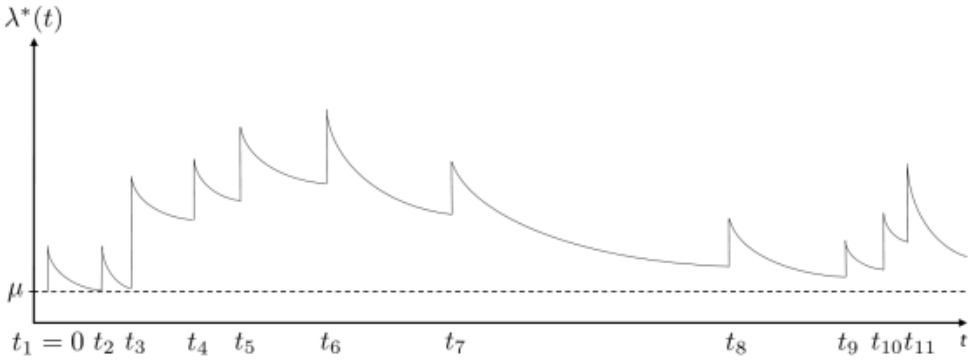
Εικόνα 3.2: Απεικόνιση της διαδικασίας Markov jump X(t)

- Μοντελοποίηση του Clickstream:

Μοντελοποιούμε το clickstream ενός χρήστη χρησιμοποιώντας μια σημειωμένη διαδικασία σημείου και επιτρέπουμε στις παραμέτρους της σημειωμένης διαδικασίας σημείου να εξαρτώνται από την λανθάνουσα φάση αγορών X(t).

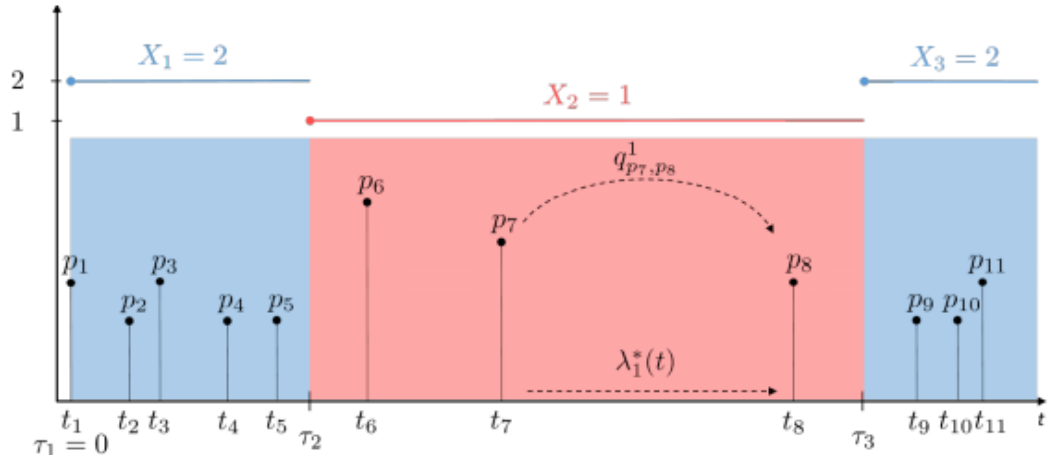
Μια σημειωμένη διαδικασία σημείου είναι μια διαδικασία συνεχούς χρόνου, όπου οι πόντοι είναι οι χρονικές περίοδοι κατά τις οποίες γίνεται επίσκεψη των σελίδων και τα σημάδια είναι οι ίδιες οι σελίδες.

$$f\left(\{\rho_{m,tm}\}_{m=1}^{m}\right) = \prod_{m} f\left(\rho_m, tm^1 H_{t_{m-1}}\right)$$



Εικόνα 3.3: Παράδειγμα απεικόνισης της πιθανοτικής αύξησης της επίσκεψης νέας σελίδας που οφείλεται στην στενή διαδοχική επίσκεψη των χρηστών.

Το μοντέλο M3PP στη συνέχεια δίνεται από το συνδυασμό της σημειακής διαδικασίας και της διαδικασίας σήμανσης, τα οποία διαμορφώνονται από την λανθάνουσα φάση αγορών processX(t). Παρακάτω στο διάγραμμα παρουσιάζεται ολόκληρο το μοντέλο για τη επίσκευψη ενός χρήστη, όπου τόσο η ένταση υπό συνθήκη όσο και η κατανομή της ακολουθίας των σελίδων που επισκέπτονται εξαρτώνται από την λανθάνουσα φάση της αγοράς.



Εικόνα 3.4: Περιγραφή του μοντέλου M3PP

4. HPM:Υβριδικό μοντέλο για την πρόβλεψη της συμπεριφοράς του χρήστη με βάση την ανάλυση N-Gram και τα αρχεία καταγραφής πρόσβασης
(HPM: A Hybrid Model for User's Behavior Prediction Based on N-Gram Parsing and Access Logs)

Το μοντέλο Markov είναι ένα δημοφιλές μαθηματικό εργαλείο που χρησιμοποιείται για την πρόβλεψη. Γενικά, ο βασικός του σκοπός είναι να προβλέπει την επόμενη ενέργεια, η οποία εξαρτάται από τα αποτελέσματα προηγούμενων. Αρκετοί ερευνητές έχουν χρησιμοποιήσει αυτήν την τεχνική με επιτυχία σε διάφορες μελέτες στη βιβλιογραφία μια από τις οποίες είναι και πρόβλεψη των καθυστερήσεων κατά την αναζήτηση στον Παγκόσμιο Ιστό.

Η συνεχής ανάπτυξη του Παγκόσμιου Ιστού έχει οδηγήσει στη δημιουργία καθυστερήσεων κατά την αναζήτηση ιστοσελίδων. Για τη μείωση των καθυστερήσεων αυτών έχουν χρησιμοποιηθεί τεχνικές "πρόβλεψης" της περιηγητικής συμπεριφοράς των χρηστών πριν την αναζήτηση των ιστοσελίδων. Το World Wide Web (-WWW) έχει γίνει σημαντικό μέρος για την κοινή χρήση πληροφοριών ο όγκος των οποίων είναι τεράστιος και αυξάνεται καθημερινά. Αρκετές μέθοδοι έχουν αναπτυχθεί την τελευταία δεκαετία , οι δύο πιο διαδεδομένες είναι η προσωρινή αποθήκευση και η προανάκτηση. [21], [22], [23]

Η εξόρυξη ιστοσελίδων μπορεί να χωριστεί σε τρείς τομείς :

1. Την εξόρυξη δεδομένων από τον Παγκόσμιο Ιστό, που περιλαμβάνει την ανάλυση προτύπων πρόσβασης χρηστών που συλλέγονται από διακομιστές web καλύτερα για την πρόβλεψη των αναγκών των χρηστών.

Το μοντέλο Markov είναι ένα μαθηματικό εργαλείο για τη στατιστική μοντελοποίηση, μία από τις δημοφιλείς μεθόδους που χρησιμοποιούνται για την πρόβλεψη. Γενικά, η βασική έννοια του μοντέλου Markov είναι να προβλέψει την επόμενη ενέργεια, η οποία εξαρτάται από τα αποτελέσματα προηγούμενων. Αρκετοί ερευνητές έχουν χρησιμοποιήσει αυτήν την τεχνική με επιτυχία σε διάφορες μελέτες στη βιβλιογραφία για να προβλέψουν τη μελλοντική τους συμπεριφορά των χρηστών.

2. Την εξόρυξη περιεχομένου web η οποία περιλαμβάνει την εξαγωγή χρήσιμων πληροφοριών από διαδικτυακούς τόπους για την εξυπηρέτηση των αναγκών των χρηστών.

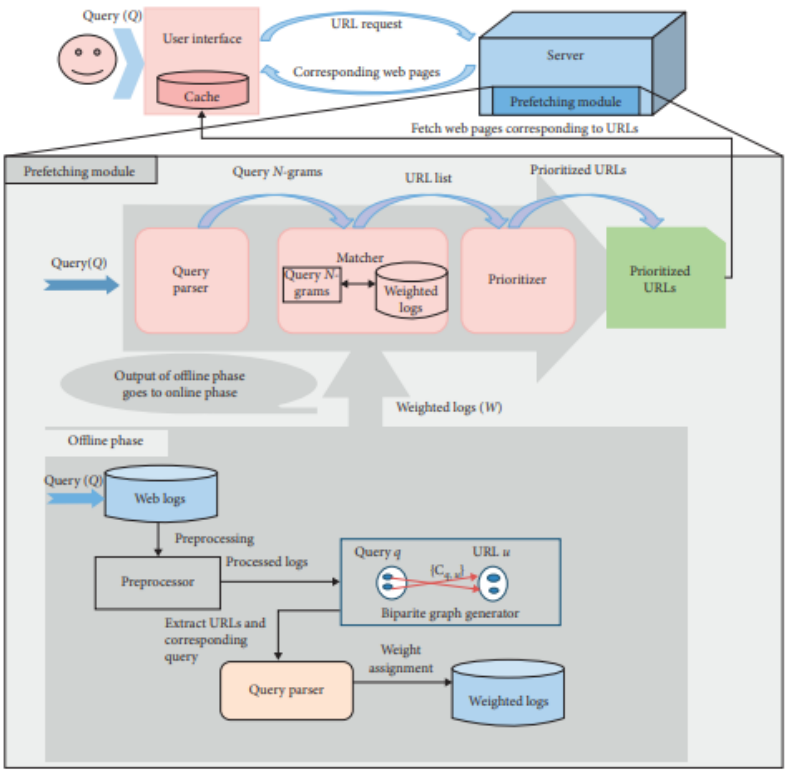
Διατηρώντας το περιεχόμενο στο επίκεντρο της ερευνητικής προσέγγισης, οι ερευνητές πρότειναν μια σημασιολογικά βελτιωμένη μέθοδο για μια ακριβέστερη πρόβλεψη που ενσωμάτωνε τις γνώσεις του τομέα και τα δεδομένα χρήσης του ιστού

του ιστότοπου. Διαπίστωσαν, όμως, ότι μόνο τα πρότυπα πρόσβασης του χρήστη είναι ανεπαρκή για να προβλέψουν τη συμπεριφορά του χρήστη.

3. Την εξόρυξη δομής που είναι η μελέτη της διασυνδεδεμένης δομής ιστοσελίδων.

Μέσω της ανάλυση της δομής των ιστοσελίδων μπορούμε να υπολογίσουμε τον τρόπο με τον οποίο οι ιστοσελίδες σχετίζονται μεταξύ τους. Οι προσεγγίσεις ανάλυσης συνδέσμων χωρίζονται σε δύο τύπους: «ανάλυση ρητής σύνδεσης» και «ανάλυση έμμεσης σύνδεσης». Οι υπερσυνδέσεις που υπάρχουν στην ιστοσελίδα ονομάζονται ρητοί σύνδεσμοι. Έχει αποδειχθεί από το Davison ότι οι πληροφορίες των υπερσυνδέσμων μπορούν να βοηθήσουν πολύ στην αναζήτηση στο διαδίκτυο. Οι σχεδιαστές ιστοσελίδων σχεδιάζουν τη δομή των συνδέσμων στην ιστοσελίδα. Οι ερευνητές χρησιμοποίησαν και τις δύο τεχνικές, και βελτίωσαν περαιτέρω την ακρίβεια αναζήτησης κατά 11,8% και 25,3%, αντίστοιχα.

Για τη βελτίωση της τεχνικής πρόβλεψης, προτείνεται ένα υβριδικό μοντέλο πρόβλεψης, το οποίο αξιοποιεί όσο το δυνατόν καλύτερα τις πληροφορίες, δηλαδή τις πληροφορίες χρήσης και τις πληροφορίες περιεχομένου των ιστοσελίδων. Το μοντέλο αυτό παρουσιάζεται σε δύο φάσεις.



Εικόνα 4.1: Μοντέλο υβριδικής πρόβλεψης

Η Offline φάση: Λειτουργεί στο παρασκήνιο και λειτουργεί περιοδικά για την ενημέρωση των καταγραφών. Δεδομένου ότι είναι ένα υβριδικό μοντέλο, η εισαγωγή

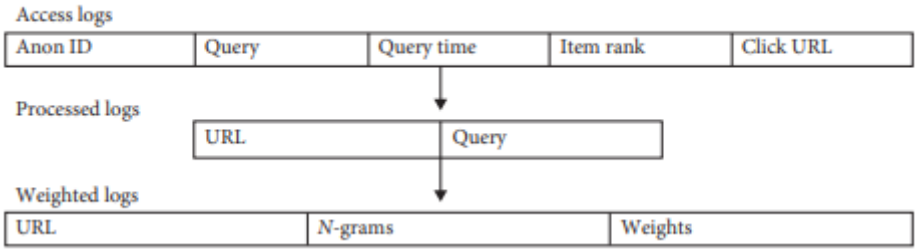
σε αυτήν τη φάση είναι τα αρχεία καταγραφής πρόσβασης και οι πληροφορίες περιεχομένου των ιστοσελίδων. Τα συνδυασμένα δεδομένα και από τις δύο πηγές τίθενται στη συνέχεια σε χρήση με διάφορα ενδιάμεσα βήματα για να γίνει μια σχετική πρόβλεψη της συμπεριφοράς των χρηστών. Το αποτέλεσμα αυτής της φάσης είναι τα σταθμισμένα αρχεία καταγραφής (WL) που περιέχουν τα N — διαγράμματα που αντιστοιχούν στα αντίστοιχα URLs.

- Προεπεξεργασία: Αρχικά, η φάση χωρίς σύνδεση εξετάζει τα αρχεία καταγραφής πρόσβασης. Τα αρχεία καταγραφής περιέχουν μια καταχώριση για κάθε αίτηση των ιστοσελίδων που γίνονται από τον πελάτη. Διάφορα πεδία των εγγραφών είναι το αναγνωριστικό ανώνυμου χρήστη, το ζητούμενο ερώτημα, η ημερομηνία και η ώρα κατά την οποία γίνεται πρόσβαση στον διακομιστή, η κατάταξη στοιχείου και η διεύθυνση URL που πατήθηκε από τον χρήστη που αντιστοιχεί στο ζητούμενο ερώτημα. Κάθε καταχώριση αρχείου καταγραφής πρόσβασης υποβάλλεται σε προεπεξεργασία για την κατάργηση των λέξεων διακοπής και την εξαγωγή του ζητούμενου ερωτήματος.
- Δημιουργία διμερούς γραφήματος: παράγεται ένα διμερές γράφημα μεταξύ των ερωτημάτων Q και των διευθύνσεων URL U. Το διμερές γράφημα επιλέχθηκε επειδή μας βοηθά να βελτιώσουμε την αναγνωσιμότητα. Αυτή η νέα αναπαράσταση γεφυρώνει φυσικά το σημασιολογικό χάσμα μεταξύ ερωτημάτων και περιεχομένου ιστοσελίδων και κωδικοποιεί πλούσιες θεματικές πληροφορίες από ερωτήματα και συμπεριφορές “κλικ” των χρηστών για την πρόβλεψη.
- Ανάλυση ερωτήματος :τα ερωτήματα που υπάρχουν στο γράφημα C αναλύονται σε N-διαγράμματα που περιγράφουν το περιεχόμενο των διευθύνσεων URL.
- Εκχώρηση βάρους: τα βάρη εκχωρούνται σε κάθε N-gram του ερωτήματος, με βάση τον αριθμό των φορών που έχει γίνει κλικ σε ένα ερώτημα, το οποίο απεικονίζεται στις άκρες από το Cq,u στο γράφημα.

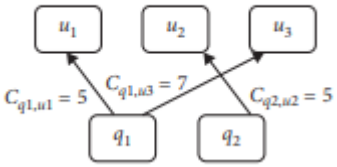
Online φάση: Όσο οι χρήστες αλληλεπιδρούν με το σύστημα, το σύστημα προβλέπει τη συμπεριφορά των χρηστών σύμφωνα με τις πληροφορίες του χρήστη. Οι πληροφορίες αυτές αντιστοιχούν με τις πληροφορίες που συλλέγονται από τα αρχεία καταγραφής της φάσης offline.

- Έναρξη ερωτήματος στη διασύνδεση: Ο χρήστης εισάγει ένα ερώτημα σύμφωνα με το ενδιαφέρον του, το οποίο μεταβαίνει στον διακομιστή μέσω μεσολαβητή χρησιμοποιώντας τη μέθοδο HTTP GET. Ο διακομιστής αποκρίνεται με τη λίστα των διευθύνσεων URL που αντιστοιχούν στο αντίστοιχο ερώτημα.

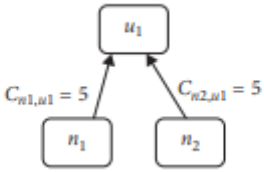
- Ενεργοποίηση αναλυτή: ενώ ο χρήστης προβάλλει την τρέχουσα σελίδα, ο διακομιστής μεσολάβησης χρησιμοποιεί αυτό το ερώτημα για περαιτέρω επεξεργασία.
- Ενεργοποίηση Matcher: αυτή η φάση λαμβάνει ως είσοδο τους όρους του ερωτήματος από την online φάση και σταθμισμένα αρχεία καταγραφής (WL) από το στάδιο εκτός σύνδεσης.
- Δημιουργία λίστας πρόβλεψης: αυτά τα βάρη στη συνέχεια μεταφέρονται στη μονάδα πρόβλεψης. Θέτει σε προτεραιότητα τις διευθύνσεις URL με βάση τα βάρη τους που δημιουργήθηκαν στο βήμα 3 και δημιουργείται μια λίστα πρόβλεψης των διευθύνσεων URL.
- Προανάκτηση: προβλέπει τις διευθύνσεις URL και τις αποθηκεύει στην cache.



Εικόνα 4.2: Οι καταγραφές που χρησιμοποιούνται στην online φάση.



Εικόνα 4.3: παράδειγμα C διαγράμματος



Εικόνα 4.4: παράδειγμα NC διαγράμματος

Τέλος, στο μέλλον οι έρευνες θα πρέπει να επικεντρωθούν στη βελτίωση του μοντέλου πρόβλεψης που αναπτύχθηκε μέχρι τώρα ώστε να ταιριάζει ακριβώς με τους όρους του ερωτήματος του ενδιαφέροντος του χρήστη. [24], [25]

Βιβλιογραφία

[1] T. P. H. E. S. a. H. S. M. C. Curme, «Quantifying the semantics of search behavior before stock market moves,» *Proceedings ofthe National Academy of Sciences,* 2014.

[2] K. Dembczynski, «Predicting Ads' Click-Through Rate with Decision Rules.,» *In Workshop on Targetingand Ranking in Online Advertising,* 2008.

[3] P. A. B. J. B. Jun B. Kim, «Online Demand Under Limited Consumer Search,» *Marketing Science 29 (6),* 2010.

[4] M. &. V. M. Eirinaki, «Web mining for Web personalization,» *ACM Transactions on Internet Technology 3,* 2003.

[5] M. T. Ö. Sule Gündüz, «A Web page prediction model based on click-stream tree representation of user behavior,» *KDD '03.,* 2003.

[6] T. H. a. S. Feuerriegel, «Early Detection of User Exits from Clickstream Data: A Markov Modulated Marked Point Process Model,» *In Proceedings of The Web Conference 2020,* 2020.

[7] S. V. J. D. N. Setia, «HPM: A Hybrid Model for User's Behavior Prediction Based on N-Gram Parsing and Access Logs,» *Sci. Program,* 2020.

[8] O. &. B. O. &. Ü. M. Can, «Personalizable Ontology-based access control,» 2010.

[9] P. &. M. A. &. J. B. &. O. T. Ruivo, «Defining the ERP and CRM Integrative Value,» *Procedia Technology,* 2014.

[10] H. Tohidi, «Why CRM is important?,» *AWERProcedia Information Technology & Computer Science,* 2012.

[11] L. Retta, «The power of collaborative filtering».

[12] W. H. N. K. K. Durno, «Expanding the boundaries of local similarity analysis,» *BMC Genomics 14,* 2013.

[13] A. M. K. Bindra, «A detailed study of clustering algorithms,» *2017 6th International Conference on Reliability, Infocom Technologies and Optimization,* 2017.

[14] J. H. M. Daniel Jurafsky, «Hidden Markov Models,» *Speech and Language Processing,* 2020.

[15] J.-P. Hosom, «Markov Model,» *International Encyclopedia of the Social & Behavioral Sciences (Second Edition),* 2003.

[16] T. Contributor, «Markov model,» *WhatIs.com,* 2017.

[17] S. R. Eddy, «Hidden Markov models,» *Current Opinion in Structural Biology,* 1996.

[18] B. J. L. Rabiner, «An introduction to hidden Markov models,» *EEE ASSP Magazine, vol. 3,* 1986.

[19] N. C. M. T. A. N. Naima Belarbi, «User Profiling in a SPOC: A method based on User Video Clickstream Analysis,» *International Journal of Emerging Technologies in Learning ,* 2019.

[20] P. J. K. J. N. Sylvain Senecal, «Consumers' decision-making process and their online shopping behavior: a clickstream analysis,» *Journal of Business Research,* 2005.

[21] C. C. J. R. K. P. M. Madhuri, « Analysis of Users' Web Navigation Behavior using GRPA with Variable Length Markov Chains,» *International Journal of Data Mining & Knowledge Management Process,* 2011.

[22] M. S. H. G. Feng W., «Markov Tree Prediction on Web Cache Prefetching,» *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing. Studies in Computational Intelligence, vol 209,* 2009.

[23] S. G. P. Temgire, «Review on Web Prefetching Techniques,» *International Journal of Technology Enhancements and Emerging Engineering Research,* 2013.

[24] M. &. G. M. &. B. G. H. Sorba, «Towards a hybrid approach for a predictive modeling of user navigational behaviors: State of the art,» *3rd International Symposium ISKO-Maghreb,* 2013.

[25] G. D. J. W. Saeed Banihashemi, «Developing a Hybrid Model of Prediction and Classification Algorithms for Building Energy Consumption,» *Energy Procedia,* 2017.