Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Τομέας Ηλεκτρικών Βιομηχανικών Διατάξεων και Συστημάτων Αποφάσεων

Εργαστήριο Διοίκησης Πληροφοριακών Συστημάτων

Αξιοποίηση Αλληλεπίδρασης του Χρήστη σε Συστήματα Διερευνητικής Αναζήτησης

Διπλωματική Εργασία

Δασκαλάκης Κωνσταντίνος

Επιβλέπων : Γρηγόρης Μέντζας

Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2014

Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Τομέας Ηλεκτρικών Βιομηχανικών Διατάξεων και Συστημάτων Αποφάσεων

Εργαστήριο Διοίκησης Πληροφοριακών Συστημάτων

Αξιοποίηση Αλληλεπίδρασης του Χρήστη σε Συστήματα Διερευνητικής Αναζήτησης

Διπλωματική Εργασία

Δασκαλάκης Κωνσταντίνος

Επιβλέπων : Γρηγόρης Μέντζας

Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 24η Οκτωβρίου 2014.

(Υπογραφή)	(Υπογραφή)	(Υπογραφή)
.....	.....	.....
Γρηγόρης Μέντζας	Ιωάννης Ψαρράς	Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2014

.....

Κωνσταντίνος Ν. Δασκαλάκης

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

Περίληψη

Σε αυτή την εργασία, το πρόβλημα που μας απασχολεί είναι η δυνατότητα ύπαρξης και υλοποίησης ενός μοντέλου που να περιγράφει την συμπεριφορά των χρηστών σε μια μηχανής διερευνητικής αναζήτησης και τι οφέλη θα μπορούσε αυτό να έχει στην λειτουργικότητα της μηχανής καθώς και στην βελτίωση εμπειρίας των χρηστών. Για να μπορέσουμε να μελετήσουμε αυτό το πρόβλημα καθώς και πως μπορούμε να εκμεταλλευτούμε τις συνήθειες των χρηστών σε μια εφαρμογή διερευνητικής αναζήτησης , χρησιμοποιήσαμε την υπάρχουσα εφαρμογή διερευνητικής αναζήτησης CRUISE (Creative User centric Inspirational Search) του εργαστηρίου IMU (Information Management Unit) του Ε.Μ.Π. Έτσι με βάση αυτή την εφαρμογή εξαγάγαμε τα δεδομένα των συνηθειών των χρηστών μέσω ιστορικού των κινήσεων τους και προσεγγίσαμε το βασικό πρόβλημα με την βοήθεια των Κρυφών Μαρκοβιανών Μοντέλων κατασκευάζοντας ένα μοντέλο που προσφέρει ικανοποιητικά αποτελέσματα προβλέποντας όσο το δυνατόν καλύτερα τις επόμενες κινήσεις των χρηστών στην εφαρμογή.

Λέξεις Κλειδιά

Αναζήτηση, Διερευνητική αναζήτηση, Αλληλεπίδραση χρηστών, Κινήσεις χρήστη, Hidden Markov Models, Πλαίσιο

Abstract

In this thesis, we will make an approach to the challenging problem of having a model of the users' behaviour while they are using an exploratory search engine, and what benefits this could have in the functionality of the engine and in the improval of the users' experience. In order to study this problem and find out how we can exploit users' habbits our study was conducted with the help of the existing exploratory search engine CRUISE (Creative User centric Inspirational Search) of the IMU lab of NTUA. Using this project we extracted all the data from the previous sessions and then used these data to create a model of their actions based on the theory of Hidden Markov Models. Our target was to build the most appropriate model for this project which can predict with high accuracy each user's next action with only feedback his past behavior.

Keywords

Search, Exploratory Search, User Interaction, User actions, Hidden Markov Models, Context

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά, τον Κ.Μέντζα που μου παραχώρησε αυτό το σημαντικό και πολύ ενδιαφέρον θέμα για να μελετήσω καθώς και την υποστήριξη του σε κάθε βήμα για το πως πρέπει να προχωρήσω σε αυτή την εργασία, καθώς και την Κα.Ταραμίγκου, που η βοήθεια της ήταν απαραίτητη, από την αρχή μέχρι και το τέλος και έπαιξε καθοριστικό παράγοντα στην σωστή διεξαγωγή όλων των δοκιμών και πειραμάτων.

Πίνακας Περιεχομένων

Ευρετήριο εικόνων

Κεφάλαιο 1

Εισαγωγή

1.1 Γενική περιγραφή εργασίας - σκοπός

Σκοπός αυτής της διπλωματικής εργασίας είναι η μελέτη και η εύρεση τρόπων αξιοποιήσης της αλληλεπίδρασης των χρηστών σε εφαρμογές μηχανών διερευνητικής αναζήτησης με σκοπό να βρούμε μοντέλα που αποτυπώνουν την συμπεριφορά τους ώστε να προκύψει κάποια πρόβλεψη για τις επόμενες κινήσεις τους. Αρχικά, θα δωθεί το απαραίτητο θεωρητικό κομμάτι για το τι είναι η διερευνητική αναζήτηση, καθώς και το πως επιδράει ο χρήστης σε αυτήν. Στο κομμάτι της υλοποίησης, χρησιμοποιήσαμε την εφαρμογή CRUISE (Creative User centric Inspirational Search) του εργαστηρίου IMU (Information Management Unit) του Ε.Μ.Π., που είναι μια εφαρμογή που παρέχει στους χρήστες δυνατότητα αναζήτησης περιεχομένου μέσων κοινωνικής δικτύωσης όπως το Twitter, εικόνες από το διαδίκτυο καθώς και συνδέσμους (links) και βίντεο. Πάνω σε αυτή την εφαρμογή βασιστήκαμε και αξιοποιήσαμε τα δεδομένα των χρηστών και των προτιμήσεων τους ανάλογα με τον τρόπο χρήσης της εφαρμογής ώστε να προκύψει αρχικά ένα μοντέλο με βάση την πιθανότητα της κάθε τους κίνησης. Τα δεδομένα αυτά προέκυψαν με την χρήση των τεχνολογιών Javascript, εξειδικευμένων βιβλιοθηκών της όπως η jQuery[1], Java καθώς και SQL ώστε να τα αποθηκεύουμε σε μια βάση δεδομένων.

Έχοντας πλέον τα δεδομένα με τα οποία με τα οποία σκοπεύουμε να μοντελοποιήσουμε τις κινήσεις των χρηστών, θα χρησιμοποιήσουμε την θεωρία των Hidden Markov Models για να το πετύχουμε αυτό. Για την υλοποίηση του μοντέλου χρειαστήκαμε την βοήθεια του υπολογιστικού εργαλείου Matlab, το οποίο μας βοήθησε για τους πολύπλοκους απαιτητικούς υπολογισμούς που χρειάστηκαν λόγω του όγκου των δεδομένων.

Τέλος μετά την συλλογή όλων των δεδομένων και την υλοποίηση του μοντέλου με τον αντίστοιχο αλγόριθμο από την θεωρία των HMM(Hidden Markov Models) , χρησιμοποιήσαμε κατάλληλες τεχνικές για να αξιολογήσουμε τα αποτελέσματα μας ως τώρα αλλά και για να κάνουμε και στην συνέχεια προβλέψεις για μελλοντικές πιθανές ακολουθίες/συνήθειες των χρηστών.

1.2 Γενικές πληροφορίες για τα εργαλεία που χρησιμοποιήσαμε στην υλοποίηση

Για να πραγματοποιήσουμε την υλοποίηση έπρεπε να χρησιμοποιήσουμε εργαλεία όπως οι γλώσσες προγραμματισμού Javascript και η βιβλιοθήκη της Jquery[1]. Επίσης, ήταν απαραίτητη η χρήση ενός συστήματος βάσεων δεδομένων ώστε να αποθηκεύουμε γρήγορα και σωστά το μεγάλο πλήθος δεδομένων που έπρεπε να χειριστούμε, και εδώ ήταν απαραίτητη η χρήση SQL και συγεκριμένα της MySql. Στην συνέχεια, βασικό κομμάτι για την εύκολη επεξεργασία των δεδομένων και υλοποίηση των αλγορίθμων ήταν η χρήση του υπολογιστικού πακέτου Matlab το οποίο διευκόλυνε κατά πολύ τους υπολογισμούς μας, που απαιτούσαν πολύ χρόνο για να ολοκληρωθούν, και φυσικά υπολογισμός χωρίς κάποια βοήθεια όπως το Matlab θα ήταν ανεύφικτος, λόγω του αριθμού των δεδομένων που έπρεπε να επεξεργαστούμε.

Στο κομμάτι της δημιουργίας και αποθήκευσης των δεδομένων των χρηστών απαραίτητη ήταν η χρήση της Javascript η οποία είναι μια ευρέως διαδεδομένη γλώσσα προγραμματισμού στο κομμάτι του διαδικτύου και οι περισσότερες ιστοσελίδες την χρησιμοποιούν για δυναμική αλληλεπίδραση των χρηστών με τις ιστοσελίδες. Το προτζεκτ CRUISE στο οποίο βασιστήκαμε για την υλοποίηση μας, όσον αφορά τη διεπαφή με το χρήστη βασίζεται στην Javascript, και συγκεκριμένα χρησιμοποιεί την βιβλιοθήκη της jQuery, άρα η χρήση της ήταν απαραίτητη .

Στο πίσω κομμάτι της εφαρμογής μας (back-end) στο οποίο μας ενδιέφερε η εγγραφή των δεδομένων στην βάση MySql, οι εγγραφές έγιναν με την βοήθεια της Java, και συγκεκριμένα με την χρήση της τεχνολογίας Hibernate Object Relational Mapping[2] η οποία παρείχε μεγάλη χρησιμότητα και ευκολία στις εγγραφές καθώς δεν υπήρχε ανάγκη χρήσης των συμβατικών εντολών της SQL αλλά μπορούσε με εντολές γραμμένες σε Java και με συγχρονισμό της εφαρμογής μας με την MySql να το επιτυγχάνει. Επίσης όλη η εφαρμογή CRUISE αναπτύχθηκε με χρήση της τεχνολογίας SPRING-MVC[3] και με την μορφή του

[1] http://jquery.com
[2] http://hibernate.org
[3] http://en.wikipedia.org/wiki/Spring_Framework,

πρότζεκτ τύπου Maven[4] το οποίο μας διευκόλυνε πολύ ώστε να μπορέσουμε να υλοποιήσουμε όλα τα ζητούμενα μας.

Τέλος, το κομμάτι της επεξεργασίας των δεδομένων, έγινε με την χρήση του Matlab όπως αναφέραμε και προηγουμένως, στο οποίο με την χρήση βιβλιοθηκών του, καθώς και άλλων βιβλιοθηκών , αλλά και προγραμμάτων που φτιάξαμε για να εξυπηρετήσουν τους σκοπούς αυτής της εργασίας, μπορέσαμε να φέρουμε εις πέρας το χρονοβόρο κομμάτι της εκπαίδευσης των μοντέλων με την χρήση των Hidden Markov Models.

1.3 Οργάνωση κειμένου

Για την διευκόλυνση του αναγνώστη, αναλύουμε πως έχει οργανωθεί το κείμενο σε όλη την εργασία, ώστε να μπορεί να ανατρέξει ευκολότερα στα σημεία που τον ενδιαφέρουν.

Άρχικά στο κεφάλαιο 1 είχαμε την εισαγωγή και λίγα λόγια για τον γενικότερο στόχο της εργασίας. Στο κεφάλαιο 2, δίνουμε μια σαφή εικόνα για το τι είναι η διερευνητική αναζήτηση, τις διαφορές από την απλή αναζήτηση καθώς και τα προβλήματα τα οποία προκύπτουν κατά την διάρκεια της.

Στην συνέχεια στο κεφάλαιο 3, βλέπουμε πως παίζει ρόλο ο χρήστης στην διερευνητική αναζήτηση και πόσο σημαντική είναι η συνεισφορά του για την υλοποίηση τέτοιων συστημάτων, καθώς και πως μπορούμε να εκμεταλλευτούμε θέματα των συνηθειών του και της ψυχολογίας του για βελτίωση του συστήματος μας.

Στο κεφάλαιο 4, αναλύουμε την δική μας προσέγγιση στο πρόβλημα, καθώς και τα συστήματα που θα χρησιμοποιήσουμε για να επιτύχουμε τα ζητούμενα μας. Δίνουμε πλήρη περιγραφή των απαραίτητων εννοιών που θα χρησιμοποιήσουμε καθώς και την θεωρία τους, ενώ στο τέλος εξηγούμε πως θα συνδυάσουμε όλες τις παραπάνω πληροφορίες.

Μετά στο κεφάλαιο 5, δίνουμε την υλοποίηση του συστήματος μας , από την αρχή μέχρι και το τέλος. Δηλαδή αναλύουμε όλες τις απαραίτητες διαδικασίες που θα ακολουθήσουμε, με τα τεχνικά τους χαρακτηριστικά.

[4] https://maven.apache.org/

Στο κεφάλαιο 6, κάνουμε την αξιολόγηση των πειραμάτων μας, και περιγράφουμε επακριβώς τα στοιχεία των τελικών περιραμάτων με αριθμούς και με στοιχεία που χρησιμοποιήσαμε. Στο κεφάλαιο 7, δίνουμε τα συμπεράσματα μας, από τα πειράματα που αναλύσαμε στο προηγούμενο κεφάλαιο, καθώς και μελλοντικούς τρόπους εκμετάλλευσης των δεδομένων αυτών.

Τέλος στο κεφάλαιο 8, έχουμε το παράρτημα της εργασίας, στο οποίο έχουμε σημαντικά κομμάτια από το πειραματικό μέρος, και συγκεκριμένα μέρη του κώδικα που χρησιμοποιήσαμε για να υλοποιήσουμε τα πειράματα και τις μετρήσεις , καθ'όλη την διάρκεια της εργασίας, καθώς και το γλωσσάρι που περιέχει μεταφράσεις όρων που χρησιμοποιήθηκαν. Μετά φυσικά ακολουθεί η βιβλιογραφία.

Κεφάλαιο 2

Διερευνητική αναζήτηση

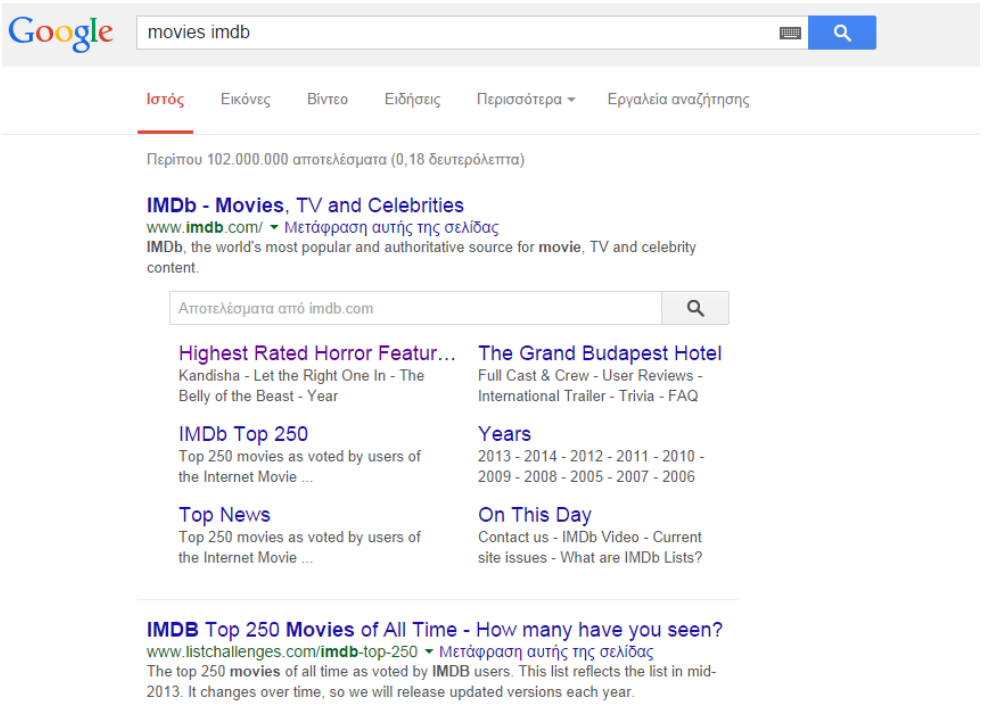
2.1 Από την απλή αναζήτηση στην διερευνητική αναζήτηση

Στην σημερινή εποχή της πληροφορίας έχει αρχίσει να γίνεται φανερό πως οι απλές αναζητήσεις που μας προσφέρουν οι γνωστές μηχανές αναζήτησεις όπως Google, Yahoo, Bing κτλ, δεν είναι αρκετές για κάποιους πιο απαιτητικούς χρήστες που αναζητούν όλο και πιο εξειδικευμένες αναζητήσεις πληροφορίας. Έτσι, η ανάπτυξη τέτοιου είδους τεχνολογιών που να προσφέρουν αυτό ακριβώς είναι αναγκαία.

Τα τελευταία χρόνια γίνεται έντονη έρευνα όσον αφορά το πρόβλημα της ανάκτησης πληροφοριών και όπως γνωρίζουμε υπάρχει μεγάλη πρόοδος σε αυτόν τον τομέα, με το πως μπορούμε να χειριζόμαστε, να αποθηκεύουμε και να ανακτούμε τις πληροφορίες είτε πρόκειται για αρχεία στον προσωπικό μας υπολογιστή, είτε πως αρχειοθετεί τα βιβλία της μια βιβλιοθήκη, είτε ένας ολόκληρος οργανισμός. Πλέον όλα αυτά γίνονται αυτόματα με την χρήση σχετικών αλγορίθμων και τεχνολογιών που έχουν την δυνατότητα να καταχωρούν και να ανακτούν όλων των ειδών τις πληροφορίες με πολύ μεγάλη ταχύτητα και πολύ αποδοτικά. Πάνω σε αυτό το μοντέλο βασίζεται ουσιαστικά και η λειτουργία των σύγχρονων μηχανών αναζήτησης του διαδικτύου.

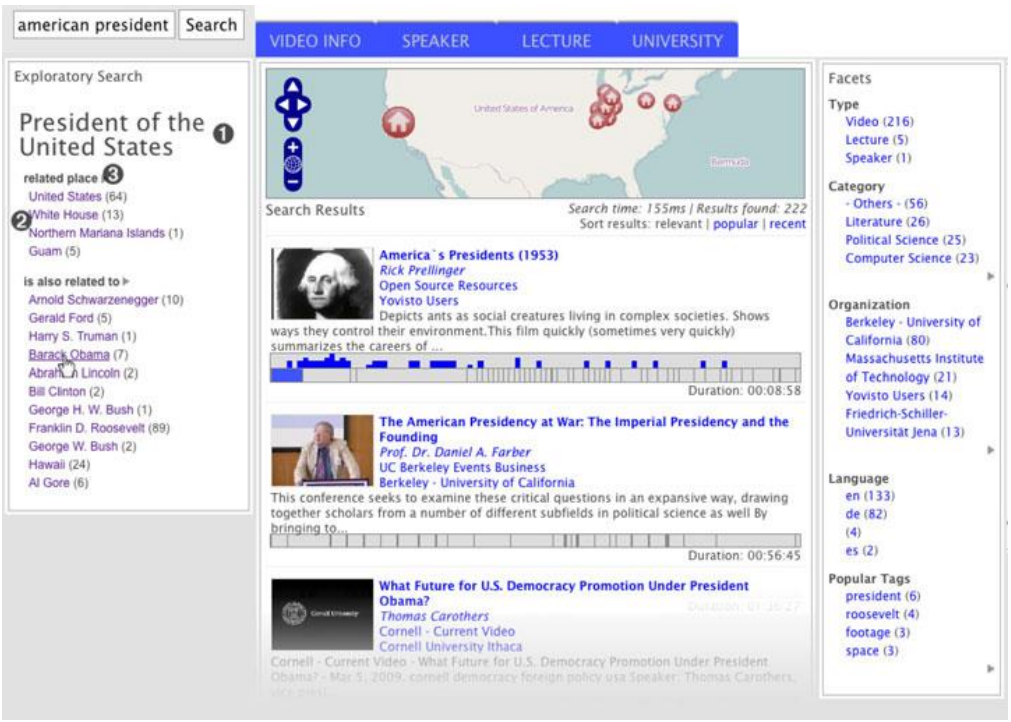
Η διερευνητική αναζήτηση (exploratory search) είναι ο τρόπος ώστε να προσπαθήσουμε να επιλύσουμε αυτό ακριβώς το πρόβλημα σε πιο σύνθετες αναζητήσεις. Δηλαδή είναι η προσέγγιση ενός προβλήματος αναζήτησης πληροφοριών το οποίο έχει πολλές όψεις, δηλαδή πολλά πιθανά αποτελέσματα αναζήτησης, και στοχεύει στο να λύνει τέτοια πολύπλοκα προβλήματα που απαιτούν συνδυαστική σκέψη, και αλληλεπίδραση ανθρώπου και πληροφορίας μέσω των υπολογιστών. Πρόσφατα μάλιστα είναι ένα αντικείμενο έρευνας που έχει γνωρίσει μεγάλη ανάπτυξη και φυσικά συνεχίζει προς αυτή την κατεύθυνση αφού αυξάνεται συνεχώς η ανάγκη για τις τεχνολογίες ανάκτησης πληροφοριών, της επιστήμης της πληροφορίας καθώς και της αλληλεπίδρασης ανθρώπου μηχανής με χρήση φυσικά ανεπτυγμένων μηχανών αναζήτησης στο διαδίκτυο. Παρακάτω

δίνονται δύο παραδείγματα μηχανών αναζήτησης, όπου η πρώτη είναι μια απλή μηχανή αναζήτησης , ενώ η επόμενη μια μηχανή διερευνητικής αναζήτησης.



Εικόνα 1 – Στιγμιότυπο μηχανής αναζήτησης

Σε αυτή την μηχανή αναζήτησης της Google, βλέπουμε πως ο χρήστης απλά κάνει αναζητήσεις με όρους, χωρίς να επεμβαίνει φιλτράροντας τα αποτελέσματα και παίρνει ακριβή αποτελέσματα με βάση τους όρους που αναζήτησε μόνο.



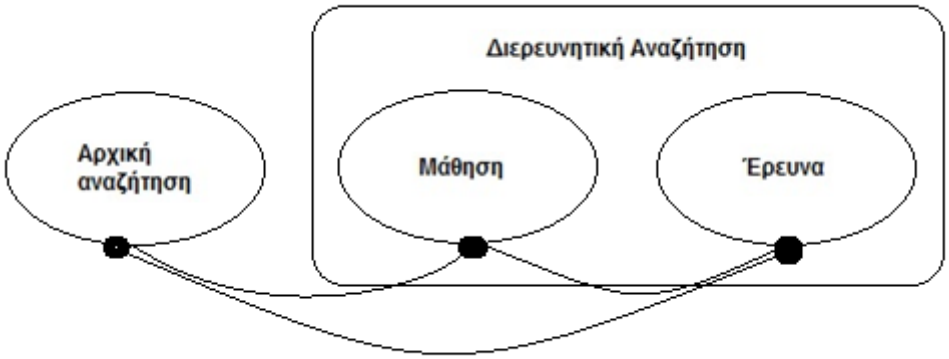
Εικόνα 2 – Στιγμιότυπο μηχανής διερευνητικής αναζήτησης

Ενώ στην δεύτερη περίπτωση βλέπουμε μια μηχανή αναζήτησης η οποία μας δίνει την δυνατότητα να φιλτράρουμε τα αποτελέσματα όπως εμείς θέλουμε ώστε να μπορούμε να κάνουμε διερευνητική αναζήτηση, ψάχνοντας όλο και πιο συγκεκριμένες πληροφορίες, σύμφωνα με τα όσα έχουμε ήδη πει για τα χαρακτηριστικά της διερευνητικής αναζήτησης. Δηλαδή ότι μια απλή μηχανή αναζήτησης όπως η πρώτη, στοχεύει στο να απαντάει με ακρίβεια ένα συγκεκριμένο ερώτημα και να δίνει αποτελέσματα με βάση την σχετικότητα τους χωρίς να παίζει άλλο ρόλο ο χρήστης σε αυτή την συγκεκριμένη αναζήτηση, ενώ στην δεύτερη περίπτωση ο χρήστης έχει ενεργό ρόλο καθώς είναι αυτός που επιλέγει από που θα προέλθουν οι πληροφορίες, και ο στόχος της είναι περισσότερο η μάθηση, η δυνατότητα επίλυσης προβλημάτων καθώς και η προβολή πιθανών λύσεων ενός προβλήματος παρά η στοχευμένη απάντηση μονοσήμαντα.

2.2 Ορισμός της διερευνητικής αναζήτησης

Ο ακριβής ορισμός της διευρευνητικής αναζήτησης είναι αρκετά πολύπλοκος και πολύπλευρος [12]. Γενικά αν προσπαθήσουμε να κατατάξουμε σε κατηγορίες τις αναζητήσεις ανάλογα το είδος τους, θα δούμε πως είναι όλες έως ένα βαθμό διερευνητικές αναζητήσεις όμως ο ακριβής ορισμός της διευρευνητικής αναζήτησης διαφέρει αρκετά αφού ουσιαστικά αποτελείται από δύο μεγάλα κομμάτια που είναι αρχικά το πρόβλημα πλαισίου δηλαδή το πρόβλημα που καλούμαστε να λύσουμε, και στην συνέχεια η ακριβής διαδικασία αναζήτησης που πρέπει να ακολουθηθεί για την επίλυση του. Δηλαδή μια αλληλουχία αναζητήσεων η οποία λύνει επιμέρους όλο και πιο εξειδιεκευμένα θέματα αποτελεί μια διερευνητική αναζήτηση. Αν κάνουμε αναζητήσεις π.χ. για να βρούμε πληροφορίες για κάποιο συγκεκριμένο θέμα για το οποίο όμως δεν γνωρίζουμε προηγουμένως άλλα πράγματα σχετικά με αυτό αρχίζουμε να αναζητούμε γενικά ότι μπορούμε να βρούμε, όπως πότε εκδόθηκε το περιοδικό, ποιος είναι ο συγγραφέας, σε ποια γλώσσα είναι κτλ. Η διαδικασία αυτή που ξεκινάμε να βρούμε αποτέλεσμα με συνεχείς αναζητήσεις και που όσο περισσότερο ψάχνουμε τόσο πιο πολύ έχουμε την δυνατότητα να φιλτράρουμε τα αποτελέσματα μας εφόσον πλέον έχουμε αρχίσει και αποκομίζουμε πληροφορίες αποτελεί ένα παράδειγμα διερευνητικής αναζήτησης. Άρα η αλληλουχία αναζητήσεων σε άγνωστο θέμα, χωρίς ακόμα να γνωρίζουμε και ποιο θα είναι το αποτέλεσμα , με παράλληλη αξιοποίηση των πληροφοριών που λαμβάνουμε σε κάθε προηγούμενη αναζήτηση αποτελεί το σύνολο που ονομάζεται διερευνητική αναζήτηση .

Φυσικά όσοι ασχολούνται με την διερευνητική αναζήτηση γνωρίζουν πως το πρόβλημα στην πράξη δεν είναι τόσο απλό ίσως όπως ακούγεται , αλλά απαιτεί συνδυασμό σωστής συμπεριφοράς αναζήτησης [1] ώστε να διασχίζουμε τις πληροφορίες με αποδοτικό τρόπο που να μας βοηθάει να επιλύουμε σωστά το πρόβλημα μας. Δηλαδή οι χρήστες 1) Δεν γνωρίζουν το ακριβές νόημα της αναζήτησης που πρόκειται να κάνουν, αλλά πρέπει πρώτα να αναζητήσουν σχετικές πληροφορίες με αυτό 2) Δεν γνωρίζουν πως θα προσεγγίσουν το πρόβλημα για να βρουν λύση σε αυτό 3) Δεν γνωρίζουν το τελικό σωστό αποτέλεσμα της αναζήτησης.



Εικόνα 3 - Βήματα διερευνητικής αναζήτησης

Και επίσης να ξεκαθαρίσουμε πως με τον όρο δερευνητική αναζήτηση όπως περιγράψαμε και νωρίτερα αναφερόμαστε σε δύο διαφορετικές διαδικασίες [1], είτε ξεχωριστά είτε μαζί. Αφενώς στο πρόβλημα πλαισίου το οποίο μας ωθεί στην αναζήτηση για την επίλυση του, και κατά δεύτερον στον τρόπο με τον οποίο διεξάγεται η αναζήτηση για την επίλυση του προβλήματος

2.3 Το πρόβλημα πλαισίου

Στο κομμάτι αυτό θα περιγράψουμε αναλυτικά τι είναι το πρόβλημα πλαισίου και πως μας επηρεάζει και εμάς σε αυτή την εργασία εφόσον η αλληλεπίδραση του χρήστη με την εφαρμογή είναι το κομμάτι πλαισίου που μας ενδιαφέρει. Γενικά το πρόβλημα πλαισίου, και η έννοια του πλαισίου είναι ένα ζήτημα που είναι αντικείμενο ενδιαφέροντος πολλών ερευνητικών πεδίων. Έτσι ανάλογα το πεδίο της εφαρμογής θα υπάρχει και διαφορετική εξήγηση του. Ένας γενικός ορισμός [11] του είναι , πως είναι η κατάσταση στην οποία

βρίσκεται ο χρήστης καθώς και οι κινήσεις του γενικότερα, ο οποίος ορισμός βρίσκει και εφαρμογή στην περίπτωση μας.

Συγκεκριμένα, ενδιαφερόμαστε για την αλληλεπίδραση πλαισίου του χρήστη (user interactional context). Συγκεκριμένα ο ορισμός πλαισίου στο θέμα μας είναι όλες οι πληροφορίες και κινήσεις με τις οποίες μπορούμε να χαρακτηρίσουμε την κατάσταση του χρήστη και τις θεωρούμε σχετικές με την αλληλεπίδραση του, με την εφαρμογή[22].

Δηλαδή θέλουμε όλες τις κινήσεις των χρηστών με τις οποίες αλληλεπιδρούν με την μηχανή αναζήτησης οι οποίες μπορεί να είναι κινήσεις όπως κλικ που κάνει σε διάφορα σημεία στην εφαρμογή ή η πληκτρολόγηση όρων. Επίσης μας ενδιαφέρει και η πνευματική κατάσταση στην οποία βρίσκεται καθώς είναι και αυτός βασικός παράγοντας για τι πρόκειται να κάνει στην συνέχεια κατά την αλληλεπίδραση του με την εφαρμογή. Ο σκοπός μας είναι να μπορέσουμε να βρούμε αυτή την σχέση ανάμεσα στις κινήσεις του χρήστη και ανάλογα στην κατάσταση που βρίσκεται κάθε στιγμή. Ο λόγος που αναφέρουμε και την πνευματική κατάσταση του είναι διότι είναι ο βασικότερος παράγοντας για τις μελλοντικές κινήσεις του και με βάση , και αντιστρόφως από τις κινήσεις να βγάλουμε συμπεράσματα και για την κατάσταση του ώστε να προβλέψουμε και την μελλοντική του κατάσταση.

2.4 Μέθοδοι αξιολόγησης συστημάτων διερευνητικής αναζήτησης

Εφόσον έχουμε πλέον αναλύσει τι ακριβώς είναι ένα σύστημα διερευνητικής αναζήτησης, ποιες διαφορές έχει με την απλή αναζήτηση καθώς και ποια προβλήματα προκύπτουν κατά την ανάλυση του όπως το πρόβλημα πλαισίου που αναλύσαμε στην προηγούμενη παράγραφο, πρέπει να δούμε πότε ένας χρήστης είναι ικανοποιημένος από ένα σύστημα διερευνητικής αναζήτησης και εκτός από αυτό να δούμε γενικότερα τι χαρακτηριστικά πρέπει να έχει , και πως μπορούμε με έναν τρόπο να αξιολογήσουμε καθολικά αυτά τα συστήματα.

Όταν αξιολογούμε συστήματα διερευνητικής αναζήτησης, είναι αδύνατον φυσικά να αφήσουμε εκτός τον ανθρώπινο παράγοντα και συγκεκριμένα την ανθρώπινη συμπεριφορά από την απόδοση του συστήματος, καθώς η αλληλεπίδραση ανθρώπου και συστήματος

είναι τόσο κοντά ώστε να μην μπορεί να διαφοροποιηθεί σε καμία περίπτωση. Όχι μόνο δεν διαφοροποιείται αλλά μπορούμε να πούμε πως συνυπάρχουν άνθρωπος και σύστημα διερευνητικής αναζήτησης σκόπιμα και λειτουργεί μάλιστα βοηθητικά ο ένας στον άλλο. Γενικά η αναζήτηση πληροφοριών είναι συνήθως συνυφασμένη και με άλλες δραστηριότητες παράλληλα, και είναι σύνηθες οι χρήστες να αναζητούν παράλληλα πληροφορίες [5] διαφορετικών θεμάτων και προβλημάτων. Και μέχρι στιγμής η περισσότερη μελέτη γίνεται στην ανάπτυξη των συστημάτων διερευνητικής αναζήτησης, και όχι τόσο στην αξιολόγηση τους, που είναι και αυτή απαραίτητη για την περαιτέρω βελτίωση τους. Μερικά πολύ σημαντικά χαρακτηριστικά [6] που θα πρέπει να συμπεριλάβουμε σε μια αξιολόγηση ενός συστήματος είναι τα παρακάτω:

Ενασχόληση και ικανοποίηση: Θέλουμε να δούμε σε ποιο βαθμό ασχολούνται οι χρήστες και εάν έχουν θετική απόκριση, δηλαδή αν νιώθουν ευχαριστημένοι κατά την διάρκεια της αναζήτησης τους, κάτι το οποίο είναι πολύ σημαντικό εργαλείο αξιολόγησης, για αυτό το λόγο βλέπουμε και πολύ συχνά σε εφαρμογές και εκτός αναζήτησης, να μας παροτρύνουν να αξιολογήσουμε την υπηρεσία τους, καθώς και σε ποιο βαθμό είμαστε ικανοποιημένοι. Φυσικά για να είναι τα αποτελέσματα αυτά αρκετά για αξιολόγηση, θα πρέπει να έχει περάσει αρκετό διάστημα ενασχόλησης ο χρήστης, καθώς και να είναι συγκεντρωμένος στην αναζήτηση του, και να μην επηρεάζεται ο βαθμός ικανοποίησης του από εξωτερικούς παράγοντες. Πέρα από την αξιολόγηση στο τέλος της κάθε συνεδρίας άλλος τρόπος αξιολόγησης είναι το αν ο χρήστης προωθεί τα αποτελέσματα του σε άλλους, ή εάν τα αποθηκεύει στα αγαπημένα του κτλ.

Καινοτομία αποτελεσμάτων: Όπως έχουμε ήδη αναλύσει, κατά την διερευνητική αναζήτηση ο σκοπός είναι να βρούμε πληροφορίες τις οποίες δεν έχουμε ξανασυναντήσει και να μας είναι εντελώς καινούριες ώστε να είμαστε κατά συνέπεια ικανοποιημένοι. Έτσι, ο βαθμός κατά τον οποίο ένας χρήστης συναντάει νέες πληροφορίες κατά την αναζήτηση του, και μετρώντας τον βαθμό αυτό μπορούμε να έχουμε ένα πολύ αποδοτικό εργαλείο αξιολόγησης των συστημάτων διερευνητικής αναζήτησης.

Βαθμός επίτευξης σκοπού: Αρχικά, να θεωρήσουμε ότι σκοπός δεν είναι μονάχα να λυθεί το αρχικό πρόβλημα του χρήστη, αλλά είναι η ικανότητα του χρήστη να βρει ένα συγκεκριμένο αποτέλεσμα που τον ενδιαφέρει [7], καθώς και να μαζέψει αρκετές χρήσιμες πληροφορίες κατά την διάρκεια της συνεδρίας του. Δηλαδή, σαν ερώτηση στον χρήστη απλά αν βρήκε αυτό που έψαχνε και σε ποιο βαθμό.

Διάρκεια: Άλλος πολύ σημαντικός παράγοντας είναι και σε πόσο χρόνο μπόρεσε ο χρήστης να ολοκληρώσει κάποιον σκοπό του αποτελεσματικά. Ο χρόνος που χρειάστηκε για να συμβεί αυτό μπορεί φυσικά να χωριστεί και σε δύο κατηγορίες, στον ωφέλιμο χρόνο που όντως βρήκε στοιχεία τα οποία και ήταν μέρος της εκπλήρωσης του στόχου, και στον χρόνο που σπατάλησε ψάχνοντας μη σχετικές πληροφορίες με το αντικείμενο μελέτης του.

Απόκτηση γνώσης: Βασικός σκοπός στην διερευνητική αναζήτηση, είναι η απόκτηση γνώσης. Το πόσα νέα πράγματα έμαθε ο χρήστης κατά την διάρκεια της αναζήτησης του είναι πολύ σημαντικός παράγοντας για την αξιολόγηση ενός συστήματος, εφόσον η διερευνητική αναζήτηση δεν στοχεύει αποκλειστικά στην επίλυση του προβλήματος αλλά και στην απόκτηση γνώσης του χρήστη.

Μέχρι στιγμής είδαμε κάποιους πολύ σημαντικούς παράγοντες με τους οποίους μπορούμε να αξιολογήσουμε τα συστήματα διερευνητικής αναζήτησης, και πως διαφέρουν πολύ σε σχέση με ένα σύστημα απλής αναζήτησης καθώς τα σημεία ικανοποίησης των χρηστών είναι τελείως διαφορετικά. Με τους τρόπους αυτούς, και με εφαρμογή τους σε διάφορες μηχανές διερευνητικής αναζήτησης μπορούμε να συγκρίνουμε τα αποτελέσματα και να δούμε ποιες είναι πληρέστερες ως προς το σύνολο τους, καθώς η αξιολόγηση σε όλα τα συστήματα είναι πάντοτε απαραίτητη για την μελλοντική βελτίωση τους.

Κεφάλαιο 3

Μοντελοποίηση και αλληλεπίδραση του χρήστη στην αναζήτηση

3.1 Μοντελοποίηση της συμπεριφοράς των χρηστών

Όπως εξηγήσαμε και προηγουμένως στην ανάλυση του προβλήματος πλαισίου, ο χρήστης παίζει καθοριστικό ρόλο κατά την διάρκεια της διερευνητικής αναζήτησης. Σε αυτό το κομμάτι θα αναλύσουμε πως είναι δυνατόν να μοντελοποιήσουμε συμπεριφορές των χρηστών ώστε να γίνουν οι εφαρμογές διερευνητικής αναζήτησης ακόμα πιο αποδοτικές.

Το σημαντικότερο ίσως κομμάτι καθ'όλη την διάρκεια της αναζήτησης είναι η ψυχολογία του χρήστη καθώς και ο βαθμός ικανοποίησης του από το κατά πόσο προχωράει αποδοτικά η αναζήτηση του. Έτσι, ενώ αρχικά έχει μεγάλο κίνητρο και λόγο να κάνει την αναζήτηση όπως εξηγήσαμε στο πρόβλημα πλαισίου, στην συνέχεια είναι πολύ πιθανόν να μειωθεί αυτό το κίνητρο. Οι περιπτώσεις που μπορεί κάτι τέτοιο να συμβεί εάν ο χρήστης δεν βρίσκει νέες χρήσιμες πληροφορίες για αυτόν , αλλά βρίσκει συνεχώς γνωστά πράγματα τα οποία δεν του προσφέρουν κάτι καινούριο. Ή επίσης μπορεί να συμβαίνει και το αντίθετο, που εμπίπτει στο πρόβλημα πλαισίου, εφόσον δεν βρίσκει καθόλου πληροφορίες και βρίσκεται σε σύγχηση. Ειδικά στα πρώτα στάδια είναι πιο συχνό αυτό το φαινόμενο, όπου ο χρήστης πρέπει να έχει την προσοχή του στραμμένη εκεί ώστε να συνεχίσει την αναζήτηση, νιώθοντας πως συνεχώς βρίσκει νέες χρήσιμες πληροφορίες για αυτόν.

Για να γίνει πιο κατανοητό θα χωρίσουμε την όλη διαδικασία αυτή σε δύο μέρη , στην διερευνητική γενική αναζήτηση, και στην συγκεκριμένη αναζήτηση [1]. Κατά την πρώτη, έχουμε το αρχικό κομμάτι αναζήτησης όπου ο χρήστης δεν γνωρίζει πολλά πράγματα για το αντικείμενο των ερευνών του, και μέσω της διερευνητικής αναζήτησης προσπαθεί να συλλέξει δεδομένα που θα τον βοηθήσουν να κάνει την έρευνα του πιο στοχευμένη. Κατά την συγκεκριμένη αναζήτηση τώρα, ο χρήστης έχει ήδη φτάσει σε κάποιο επιθυμητό επίπεδο γνώσης για το θέμα, και θέλει πλέον να ψάξει σε βάθος ώστε να βρει περισσότερα στοιχεία για αυτό. Φυσικά, και οι δύο τρόποι είναι απαραίτητοι, και πρέπει να εμπεριέχονται σωστά.

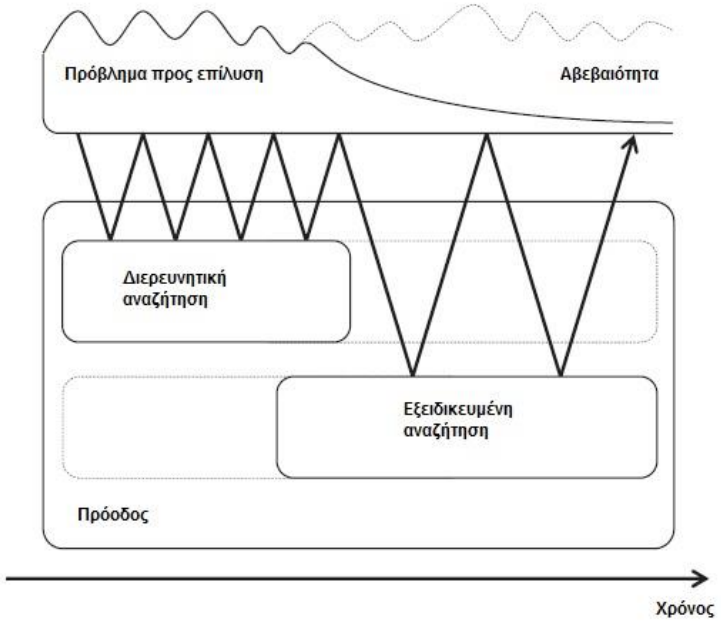
Όσον αφορά τώρα την διερευνητική αναζήτηση, η οποία προκύπτει πρώτη, ο χρήστης για να μπορέσει να ψάξει αποδοτικά θα πρέπει να κάνει κάποιες ενέργειες πρώτα πέρα από την απλή ανάγνωση. Θα πρέπει να ψάξει συγκεκριμένα και να ορίσει με ακρίβεια τις πληροφορίες που χρειάζεται. Όπως και στο διαδίκτυο, που όταν ξεκινάμε μια αναζήτηση, καθοδηγούμαστε από σελίδα σε σελίδα μέσω των συνδέσμων, ώστε να μαζέψουμε τις αρχικές πληροφορίες για να αρχίσουμε την έρευνα μας. Ο χρήστης δηλαδή κατά την πρώτη αυτή διαδικασία, θα περιηγηθεί ανάμεσα σε πλήθος πληροφοριών και θα φιλτράρει τις χρήσιμες από τις άχρηστες, όλο αυτό βέβαια αξιοποιώντας τις γνώσεις του, και με την χρήση υπόθεσης εφόσον δεν μπορεί να είναι σίγουρος εκ των προτέρων. Όπου πιθανόν είναι να επιστρέψει ακόμα και στις ίδιες πληροφορίες στην συνέχεια, αν από λάθος έχει απορρίψει κάποιες από αυτές στα πρώτα στάδια όπου δεν γνώριζε καλά το θέμα. Αφού ο χρήστης έχει πλέον ολοκληρώσει το κομμάτι της αρχικής διερευνητικής αναζήτησης , θα θελήσει στην συνέχεια να προχωρήσει στην συγκεκριμένη αναζήτηση, σύμφωνα με τις πληροφορίες που έχει μαζέψει έως τώρα. Επίσης, να ξεκαθαρίσουμε πως αυτή η διαδικασία μοιάζει κατά πολύ ανάμεσα στους χρήστες και οι συνήθειες τους είναι οι ίδιες, και μπορούν να παρατηρηθούν ομοιότητες μεταξύ τους, στην περίπτωση που μοιράζονται την ίδια εφαρμογή αναζήτησης.

Προχωρώντας τώρα στην συγκεκριμένη αναζήτηση, ο χρήστης δεν είναι πλέον απλός αναγνώστης των όσων βρίσκει, [1] αλλά πλέον αξιολογεί τα αποτελέσματα και ελέγχει αν καλύπτουν τις ανάγκες του. Έχει πλέον ξεκαθαρίσει την κατάσταση στο μυαλό του, όσον αφορά τουλάχιστον κάποιο συγκεκριμένο υποπρόβλημα, και ελέγχει εάν τα αποτελέσματα που παίρνει συμβαδίζουν με αυτά που θέλει να βρει, και αν η διαδικασία της διερευνητικής αναζήτησης που ακολούθησε προηγουμένως είναι ικανοποιητική. Εδώ η μηχανή αναζήτησης χρησιμοποιείται για αυτό μόνο το λόγο, δηλαδή την αλληλεπίδραση του χρήστη μετά την αρχική αναζήτηση. Αυτό το κομμάτι της αλληλεπίδρασης του χρήστη με τη μηχανή αναζήτησης, είναι που μπορούμε να κρατήσουμε με την αποθήκευση των κινήσεων του κατά την κάθε του συνεδρία, και θα το δούμε αναλυτικά πως γίνεται στο παρακάτω κομμάτι, αφού αυτό ακριβώς κάνουμε κρατώντας τις κινήσεις των χρηστών του CRUISE.

Ένας ακόμα λόγος, αυτού του κομματιού αναζήτησης είναι ότι οι χρήστες δεν είναι σε κάποιες περιπτώσεις βέβαιοι για τα αποτελέσματα τους, όσον αφορά την ακρίβεια που έχουν. Έτσι για να πετύχουν μεγαλύτερη ακριβεια ψάχνουν για το ίδιο πράγμα αναλυτικά σε περισσότερες των μια πηγές. Βέβαια αυτό μπορεί να οδηγήσει στην συνέχεια σε σύγχηση των χρηστών εφόσον κατά την αναζήτηση αυτή θα βρεθούν αντιμέτωποι με

μεγάλο πλήθος μη σχετικών αποτελεσμάτων. Για αυτό το λόγο και είναι σημαντικό να κρατάμε αυτά τα δεδομένα κατά τις συνεδρίες των χρηστών και κατά την αξιολόγηση τους να μπορέσουμε να τα αναλύσουμε και να βρούμε τρόπους βελτίωσης τους, όπως με την προβολή μεμονωμένων αποτελεσμάτων στους χρήστες σε αυτό το σημείο.

Ανάλογα το μέγεθος και την δυσκολία του προβλήματος, αυτή η διαδικασία μπορεί να επαναλαμβάνεται και να συμβαίνει ανάμεσα σε διάφορες συνεδρίες του χρήστη. Αυτό είναι λογικό καθώς ο χρήστης πιθανών να θελήσει να συνεχίσει την αναζήτηση κάποια άλλη χρονική στιγμή. Και για αυτό το λόγο πρέπει να είμαστε σε θέση να αναγνωρίζουμε αυτές τις περιπτώσεις ώστε τα αποτελέσματα με τα οποία τον τροφοδοτούμε να είναι σχετικά και περιορισμένα όπως τα θέλει. Προκύπτει δηλαδή από όλα αυτά, πόσο σημαντικό είναι να μοντελοποιήσουμε τις κινήσεις του χρήστη ώστε να καταλαβαίνουμε από τις κινήσεις του, σε ποια κατάσταση αναζήτησης βρίσκεται και να προσαρμοζόμαστε κατάλληλα στις ανάγκες του. Και είναι οι κινήσεις του χρήστη το μοναδικό εργαλείο που έχουμε για να μας βοηθήσει κατά το έργο αυτό. Είναι δυνατόν βέβαια να υπάρξει και άμεση εναλλαγή μεταξύ των δύο αυτών καταστάσεων, κάτι το οποίο δυσκολεύει το έργο μια ιδανικής μηχανής διερευνητικής αναζήτησης, αλλά και πάλι , όσο ακριβέστερο είναι το μοντέλο που έχουμε δημιουργήσει και περιγράψει τον χρήστη τόσο καλύτερα θα είναι τα αποτελέσματα, και η χρηστικότητα της εφαρμογής.
Κάπως έτσι θα μπορούσαμε να περιγράψουμε τις καταστάσεις από τις οποίες περνάει ο χρήστης κατά την διάρκεια της αναζήτησης, όπως φαίνονται στο παρακάτω σχήμα.



Εικόνα 4- Εναλλαγή καταστάσεων του χρήστη κατά την διάρκεια της αναζήτησης (βασισμένο στο [1])

Συγκεκριμένα βλέπουμε στο παραπάνω σχήμα το μοντέλο συμπεριφοράς των χρηστών κατά τα διάφορα στάδια της διερευνητικής αναζήτησης που έχουμε περιγράψει ως τώρα. Δηλαδή την διαδικασία στο μυαλό του χρήστη από την έναρξη της αναζήτησης και πως σκέφτεται και περνάει από το ένα στάδιο στο άλλο με την πάροδο του χρόνου, και πως η διαδικασία αυτή μπορεί να επαναλαμβάνεται πολλές φορές ακόμα και σε μία συνεδρία. Η ανάλυση της παραπάνω διαδικασίας, και όσο περισσότερες τέτοιες συνεδρίες έχουμε, τόσο καλύτερα εκπαιδεύουμε το μοντέλο, ώστε να μπορεί στην συνέχεια να αντιλαμβάνεται το στάδιο που βρισκόμαστε όσο το δυνατόν καλύτερα. Επίσης αξίζει να σημειώσουμε πως η στιγμή που ο χρήστης περνάει από την πρώτη φάση της διερευνητικής αναζήτησης στην εξειδικευμένη αναζήτηση δεν είναι βέβαιο πως βρήκε αυτό που έψαχνε, καθώς υπάρχει περίπτωση, απλά να ψάχνει περισσότερο κάποιο συγκεκριμένο θέμα για να αξιολογήσει αν όντως είναι χρήσιμο και έτσι να επιστρέψει πίσω στο πρώτο στάδιο. Όπως καταλαβαίνουμε αυτό δυσκολεύει και άλλο την ανάπτυξη του βέλτιστα εκπαιδευμένου μοντέλου, καθώς η διαδικασία αυτή δεν είναι μονοσήμαντη.

3.2 Είδος αναζήτησης ανάλογα με το αρχικό κίνητρο του χρήστη

Όπως αναφέραμε και κατά την ανάλυση του προβλήματος πλαισίου, ένα πολύ καθοριστικό κομμάτι της διερευνητικής αναζήτησης είναι ο αρχικός λόγος που ο χρήστης κάνει την αναζήτηση, και εδώ θα δούμε πως είναι δυνατόν ανάλογα το κίνητρο να οδηγείται και σε διαφορετικού είδους αναζήτηση, όσον αφορά το αν είναι αρχική διερευνητική αναζήτηση είτε συγκεκριμένη αναζήτηση, με βάση τα όσα εξηγήσαμε στην παραγραφο 3.1.

Αν πάρουμε ως κίνητρο του χρήστη που θέλει να κάνει την αναζήτηση, την περιέργεια του, για να μάθει πληροφορίες για ένα θέμα που τον ενδιαφέρει. Η περιέργεια του μπορεί να κατηγοριοποιηθεί με δύο τρόπους. Αφενώς μπορεί να είναι περιέργεια για κάτι πολύ συγκεκριμένο, ή να είναι γενική περιέργεια για κάποιο αφηρημένο θέμα.

Εάν η περιέργεια του είναι συγκεκριμένη, τότε εννοούμε πως ενδιαφέρεται να μάθει πολύ συγκεκριμένες πληροφορίες ώστε να μπορέσει να λύσει κάποιο πρόβλημα που θέλει και έχει πολύ συγκεκριμένους στόχους. Εάν τώρα η περιέργεια του είναι γενικότερου είδους τότε μπορεί να είναι απλό ενδιαφέρον για κάποιο γενικό θέμα, ή απλά να κοιτάζει πληροφορίες χωρίς να έχει κάτι συγκεκριμένο που ψάχνει [13]. Όπως γίνεται κατανοητό, κάτι τέτοιο οδηγεί σε διαφορετική χρήση της μηχανής αναζήτησης , ακόμα και σε αυτήν την

περίπτωση που θα μπορούσαμε να πούμε πως το κίνητρο είναι κοινό δηλαδή η περιέργεια.

Δηλαδή κατά την πρώτη περίπτωση, τα βήματα στο μυαλό του χρήστη είναι τα εξής. Ξεκινώντας από συγκεκριμένη περιέργεια για ένα θέμα, αρχίζοντας την αναζήτηση ο σκοπός του γίνεται αμέσως πιο συγκεκριμένος, με αποτέλεσμα να οδηγείται σε συγκεκριμένη αναζήτηση στο τέλος. Από την άλλη μεριά , αν η περιέργεια είναι γενική, αρχίζοντας την αναζήτηση βλέπει πως δεν έχει ξεκάθαρους σκοπούς , με αποτέλεσμα στο τέλος η αναζήτηση του να γίνεται γενική διερευνητική αναζήτηση. Φυσικά όλο αυτό έχει να κάνει και με την ψυχολογία του κάθε χρήστη ο οποίος ανάλογα τις συνήθειες του λειτουργεί και διαφορετικά. Δηλαδή ακόμα και ο ίδιος χρήστης, μπορεί να αλλάξει γνώμη και η περιέργεια του να στοχοποιηθεί κάπου κατά την αρχική αναζήτηση και να οδηγηθεί σε συγκεκριμένη αναζήτηση. Αυτό που γίνεται εμφανές όμως από αυτό το παράδειγμα είναι πως είναι απαραίτητο να κατανοήσουμε τα μοτίβα αυτά ώστε να μπορεί η μηχανή διερευνητικής αναζήτησης να ανταπερξέρχεται κατά τον βέλτιστο τρόπο, και να περνάει και αυτή επιτυχώς από την μία κατάσταση του χρήστη στην άλλη , άμεσα και χωρίς ιδιαίτερα στοιχεία όπως κάνει και ο χρήστης.

3.3 Αλληλεπίδραση του χρήστη και σημασία κατανόησης των προθέσεων του

Όπως γίνεται κατανοητό από τα παραπάνω ο χρήστης παίζει καίριο ρόλο κατά την αναζήτηση, και συγκεκριμένα ο σκοπός αναζήτησης του χρήστη, και έχει τραβήξει μάλιστα πολύ ενδιαφέρον η συγκεκριμένη ερευνητική περιοχή. Αν δηλαδή οι μηχανές αναζήτησης μπορέσουν να αντιλαμβάνονται τον σκοπό του χρήστη κάθε φορά, αυτό θα είχε μεγάλο αντίκτυπο σε όλες τις εφαρμογές του διαδικτύου, και θα επέτρεπε τρομερή βελτίωση τους όσον αφορά την εμπειρία χρήσης και λειτουργικότητας τους. Αυτό όπως είπαμε μπορεί να επιτευχθεί αποθηκεύοντας και αναλύοντας την αλληλεπίδραση του χρήστη με την μηχανή αναζήτησης σύμφωνα με τις αναζητήσεις που κάνει, με βάση όσα αναλύσαμε στο κεφάλαιο αυτό για τον χρήστη και την ψυχολογία του καθ’όλη την διάρκεια της αναζήτησης. Μέχρι σήμερα η πλειοψηφία των συστημάτων αυτών αγνοούν εντελώς τις προθέσεις του χρήστη σύμφωνα με τις προηγούμενες του αναζητήσεις, όμως πληθώρα μελετών δείχνουν πως αυτά τα δεδομένα μπορούν να χρησιμοποιηθούν με μεγάλη επιτυχία, και να μοντελοποιηθούν έτσι ώστε να παράγουν καλύτερα αποτελέσματα στους χρήστες.

Η επιτυχία των μηχανών αναζήτησης αλλά και άλλων εφραμογών του διαδικτύου που χρησιμοποιούν εν μέρει μηχανές αναζήτησης για την περιήγηση μας, βασίζεται στην ικανότητα των μηχανών να δίνουν στους χρήστες τα αποτελέσματα που ψάχνουν συγκεκριμένα, χωρίς να κουράζουν με πλήθος άχρηστων πληροφοριών. Φυσικά αυτό δεν είναι καθόλου βέβαιο και οι μηχανές πρέπει να προσαρμόζονται στον εκάστοτε χρήστη για να το πετύχουν όσο το δυνατόν καλύτερα. Για αυτό και καταλαβαίνουμε πόσο σημαντικό είναι για να επιτύχει μια μηχανή αναζήτησης να ενσωματώνει στοιχεία από τις προθέσεις του χρήστη.

Ένα απλό παράδειγμα για να δείξουμε το θέμα πρόθεσης του χρήστη, είναι όταν ξεκινάμε να κάνουμε μια αναζήτηση, πως πριν προλάβουμε να ολοκληρώσουμε τον όρο που θέλουμε να ανζητήσουμε, ήδη αυτός εμφανίζεται για εμάς και μας γλιτώνει από χρόνο. Αυτό βέβαια είναι σε επίπεδο γλωσσικό, όπου οι μηχανές έχουν καταχωρημένες λέξεις και μας εμφανίζουν αποτελέσματα βάση αποκλεισμού άλλων. Αυτό όμως είναι που θέλουμε να επιτευχθεί σε όλα τα επίπεδα μιας μηχανής αναζήτησης, προτείνοντας μας εκείνη, άλλες σχετικές αναζητήσεις, και γνωρίζοντας ποια θα είναι η επόμενη μας αναζήτηση, φυσικά χωρίς να ενοχλεί τον χρήστη καθώς θα πρέπει να είναι και διακριτική. Η έρευνα σε αυτό τον τομέα μάλιστα προχωράει με ραγδαίους ρυθμούς και όλο και περισσότερες εφαρμογές του διαδικτύου δείχνουν ενδιαφέρον στο να επιτύχουν πρόβλεψη των προθέσεων των χρηστών κρατώντας πλήρες ιστορικό των κινήσεων του.

Φυσικά για να συμβεί αυτό χρειάζεται εκτενής ανάλυση των κινήσεων του, και δημιουργία ενός μοντέλου που θα κάνει αυτή τη δουλειά σωστά, όπως και θα δούμε στην συνέχεια στο πειραματικό κομμάτι της εργασίας. Πρέπει επίσης να αναφέρουμε πως τα στοιχεία τα οποία πρέπει να συλλέξουμε για τους χρήστες δεν μπορεί να είναι γενικά δεδομένα αλλά θα πρέπει να είναι εξειδικευμένα και φιλτραρισμένα ώστε να μπορούν να κατηγοριοποιηθούν στη συνέχεια πριν χρησιμοποιηθούν για την εκπαίδευση του μοντέλου. Έτσι από τις κινήσεις ενός χρήστη σε μια μηχανή αναζήτησης καλό θα ήταν αρχικά να χωρίσουμε τις αναζητήσεις του σε τρεις διαφορετικές κατηγορίες. Στις αναζητήσεις οι οποίες είναι καθαρά για εύρεση πληροφοριών, σε αυτές που χρησιμεύουν για να βρει άλλες πληροφορίες, δηλαδή κατευθυντήριες αναζητήσεις, και τέλος μεταβατικές αναζητήσεις. Ανάλογα τώρα την κατηγορία που προέρχονται τα δεδομένα από την αλληλεπίδραση του χρήστη, διαφορετική επεξεργασία θα προκύψει και άλλοι αλγόριθμοι θα πρέπει να χρησιμοποιηθούν . Για να μπορέσουμε τώρα να κατανοήσουμε πως θα δημιουργηθεί ένα μοντέλο με βάση την αλληλεπίδραση του χρήστη, είναι βασικό να

χωρίσουμε την διαδικασία σε τρία κομμάτια, στην συνεδρία του χρήστη, στις αναζητήσεις κατά την συνεδρία του, καθώς και στην αλληλεπίδραση του.

Παίρνοντας όλα αυτά τα δεδομένα θα μπορέσουμε στην συνέχεια να τα εκμεταλλευτούμε για να βγει το τελικό μοντέλο. Ως συνδερία εννοούμε την αλληλουχία αναζητήσεων του χρήστη που σκοπεύουν στην επιτυχία ενός στόχου. Εάν δηλαδή ο χρήστης δεν βρει το επιθυμητό αποτέλεσμα κατά την πρώτη αναζήτηση, θα προσπαθήσει σταδιακά να επιτύχει τον σκοπό του αλλάζοντας τους όρους της αναζήτησης του. Η διαδικασία αυτή χαρακτηρίζεται ως μια συνεδρία του χρήστη με την μηχανή αναζήτησης. Στην συνέχεια, έχουμε τον όρο της αναζήτησης, που αποτελέιται από έναν συνδιασμό κινήσεων του χρήστη, που είναι η υποβολή μιας αναζήτησης, η ανάλυση των αποτελεσμάτων του καθώς και η περιήγηση στα αποτελέσματα που έχουν προκύψει. Και σε αυτή την περίπτωση ο χρήστης έχει έναν σαφή στόχο όπως και κατά την διάρκεια της συνεδρίας. Τέλος, έχουμε την περιήγηση του χρήστη σε μια ιστοσελίδα και περιλαμβάνει όλες τις κινήσεις του σε αυτήν όπως το κλικ του κέρσορα του ποντικιού, την αναζήτηση στη σελίδα πηγαίνοντας πάνω κάτω,τι γράφει στην σελίδα , και γενικά οποιαδήποτε αλληλεπίδραση μπορεί να έχει, ακόμα και τον ενεργό χρόνο που ξόδεψε κτλ. Αυτό ακριβώς το κομμάτι θα υλοποιηθεί και στο επόμενο βήμα στο κεφάλαιο 4. Όλα αυτά τα παραπάνω στοιχεία πρέπει να ληφθούν υπόψιν κατά την διαδικασία της μοντελοποίησης της αλληλεπίδρασης του χρήστη με την μηχανή αναζήτησης.

Κεφάλαιο 4

Η προσέγγιση μας – Ανάλυση των μεθόδων που θα χρησιμοποιήσουμε

4.1 Θεωρία Hidden Markov Models

Γενικά ένα Hidden Markov Model, εν συντομία και ως HMM, αποτελεί έναν τύπο στοχαστικής μοντελοποίησης, το οποίο και είναι κατάλληλο για μη στάσιμες στοχαστικές ακολουθίες , οι οποίες χαρακτηρίζονται από στατιστικές ιδιότητες με τυχαίες μεταβάσεις μεταξύ k διαφορετικών στάσιμων διεργασιών. Δηλαδή, με τα HMM μπορούμε να μοντελοποιήσουμε διεργασίες οι οποίες είναι τμηματικά στάσιμες(δηλαδή οι στατιστικές τους ιδιότητες δεν μεταβάλλονται καθώς εξελίσσεται ο χρόνος). Αν έχουμε για παράδειγμα μια ακολουθία από n παρατηρήσεις οι οποίες μπορεί και να έχουν προκύψει και από διαφορετικές πηγές, μπορούμε να την χρησιμοποιήσουμε και να παράγουμε δεδομένα με ακολουθιακό τρόπο από κάποια στατιστική κατανομή όπως την Gaussian. Άρα με τα HMM μια τέτοια στάσιμη ακολουθία για την οποία και δεν γνωρίζουμε από που πηγάζουν οι πληροφορίες μπορούμε να την μοντελοποιήσουμε.

Η μοντελοποίηση τώρα είναι πολύ σημαντικό κομμάτι στην κατασκευή των HMM. Αφενός πρέπει να θεωρήσουμε τον αριθμό των πηγών (k) γνωστό για να προχωρήσουμε σε υπολογισμούς. Στην συνέχεια θα πρέπει να προσδιορίσουμε τις πυκνότητες πιθανότητας που θα περιγράφουν τις καταστάσεις που έχουμε p(x|j),j=1,….,k. Επίσης, χρειαζόμαστε και τις πιθανότητες μετάβασης καταστάσεων P(i|j),i=1,….,k με πιθανότητα P(i|j) την πιθανότητα για μετάβαση από την κατάσταση j στην κατάσταση i. Και επειδή πρέπει να δούμε και αρχικά πως θα αρχίσουμε , θα έχουμε και μια ακολουθία για τις αρχικές τιμές, δηλαδή την P(i),i=1,…,k. Φυσικά αν η ακολουθία μας αποτελείται από διακριτές τιμές τότε οι πυκνότητες πιθανότητες θα αποτελούν πιθανότητες.

Άρα ουσιατικά η περιγραφή ενός HMM αποτελείται από:

1) τον αριθμό των αρχικών καταστάσεων , έστω k.
2) Τις πυκνότητες πιθανότητας p(x|j),j=1,….,k, όπου x=1,….,L
3) Τον πίνακα μετάβασης καταστάσεων

4) Το διάνυσμα των αρχικών πιθανοτήτων

Ο πίνακας μετάβασης καταστάσεων θα ονομάζεται πίνακας A (state transition matrix) και θα είναι τετραγωνικός μεγέθους k.

$$A = \begin{bmatrix} P(1/1) & P(2/1) & \ldots\ldots & P(k/1) \\ \ldots & \ldots & & \ldots\ldots \\ P(1/k) & P(2/k) & \ldots\ldots & P(k/k) \end{bmatrix}$$

Ο πίνακας πυκνότητας πιθανότητας θα είναι ο πίνακας B (observation probability matrix) και θα έχει διαστάσεις k x L.

$$B = \begin{bmatrix} P(x=1/1) & P(x=1/2) & \ldots\ldots & P(x=1/k) \\ \ldots & \ldots & & \ldots\ldots \\ P(x=L/1) & P(x=L/2) & \ldots\ldots & P(x=L/k) \end{bmatrix}$$

Ο πίνακας π των αρχικών πιθανοτήτων,

$$\pi = \begin{bmatrix} P(1) \\ P(2) \\ .. \\ .. \\ P(k) \end{bmatrix}$$

Εκτός βέβαια από το κομμάτι της μοντελοποίησης εξίσου σημαντικό είναι και της αναγνώρισης και εκπαίδευσης. Στο κομμάτι της αναγνώρισης υποθέτουμε πως έχουμε πάνω από ένα HMM διαθέσιμα με διαφορετικό σύνολο παραμέτρων το καθένα και διαφορετικών πηγών. Κατά την διαδικασία της αναγνώρισης ο στόχος είναι να βρούμε το μοντέλο HMM το οποίο έχει την μεγαλύτερη πιθανότητα να έχει εκπέμψει μια συγκεκριμένη ακολουθία. Συγκεκριμένα οι αλγόριθμοι που χρησιμοποιούνται για αυτό το μέρος της αναγνώρισης είναι η μέθοδος Baum-Welch ή η μέθοδος Viterbi. Και οι δύο μέθοδοι υπολογίζουν για κάθε HMM τις πιθανότητες να προκύψει μια ακολουθία, και κρατούν αυτό με την μεγαλύτερη πιθανότητα. Για να φτάσουμε όμως στο κομμάτι της αναγνώρισης και εκπαίδευσης θα πρέπει πρώτα να έχουμε εκτιμήσει όλες τις επιμέρους παραμέτρους του μοντέλου. Τέλος, στο κομμάτι της εκπαίδευσης γίνεται εκτίμηση των παραμέτρων. Συγκεκριμένα, χρειαζόμαστε ακολουθίες παρατηρήσεων ικανοποιητικού μήκους που έχουν προκύψει από αυτήν την διεργασία και με βάση αυτών να εκτιμήσουμε την πιθανότητα να προκύψουν. Για την διαδικασία αυτή πάλι χρησιμοποιούνται οι αλγόριθμοι Baum-Welch και Viterbi, αλλά και ο Forward.

Για να μπορέσει να γίνει η διαδικασία χρήσης των μοντέλων που θα ακολουθήσουμε πιο κατανοητή θα δώσουμε και ένα μικρό παράδειγμα χρήσης τους, και των αποτελεσμάτων που δίνουν. Στον πίνακα Α έχουμε τις πιθανότητες μετάβασης από την μία κατάσταση στην αλλη. Δηλαδή με βάση τον παρακάτω πίνακα Α:

$$\begin{array}{c} \\ H \\ C \end{array} \begin{array}{c} \begin{array}{cc} H & C \end{array} \\ \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} \end{array}$$

Βλέπουμε πως οι πιθανότητες που βλέπουμε σημαίνουν πως, για να παραμείνει στην κατάσταση Η από την Η έχουμε 0.7, ενώ για να πάμε από την Η στην C έχουμε πιθανότητα 0.3. Ομοίως από την κατάσταση C για να πάμε στην κατάσταση Η έχουμε πιθανότητα 0.4, ενώ για να παραμείνουμε στην C 0.4 . Επίσης, έστω ότι έχουμε και τον παρακάτω πίνακα πυκνότητας πιθανότητας Β.

$$\begin{array}{c} \\ H \\ C \end{array} \begin{array}{c} \begin{array}{ccc} S & M & L \end{array} \\ \begin{bmatrix} 0.1 & 0.4 & 0.5 \\ 0.7 & 0.2 & 0.1 \end{bmatrix} \end{array}$$

Ο οποίος υποδεικνύει πως αν είμαστε στην κατάσταση Η, οι πιθανότητες για να έχουμε τις συνθήκες S , M, L είναι 0.1 , 0.4 και 0.5 αντίστοιχα, και αν είμαστε στην κατάσταση C οι πιθανότητες για να έχουμε τις συνθήκες S , M, L είναι 0.7 , 0.2 και 0.1.

Το παραπάνω μοντέλο που περιγράψαμε αποτελείται από τους δύο αυτούς πίνακες, και αυτό θα είναι το αρχικό μας ζητούμενο, να παράξουμε αυτό το μοντέλο από τα δεδομένα που έχουμε συλλέξει ως τώρα. Το παράδειγμα που δείξαμε είναι ένα μοντέλο $1^{ης}$ τάξης εφόσον οι πιθανές καταστάσεις που έχουμε θα είναι μονάχα δύο και αφού η πιθανότητα να προκύψει η επόμενη κατάσταση εξαρτάται μονάχα από την προηγούμενη με βάση τον πίνακα Α. Όμως οι πραγματικές καταστάσεις του μοντέλου είναι άγνωστες και δεν γνωρίζουμε τίποτα για τις παρελθοντικές καταστάσεις του μοντέλου, δηλαδή είναι κρυφές οι συνθήκες, εξού και το όνομα Hidden Markov Models, και η μοναδική πληροφορία που έχουμε είναι ο πίνακας Β που μας δίνει την πυκνότητα πιθανοτήτων και από τον οποία θα πάρουμε τις πληροφορίες που θέλουμε. Να προσθέσουμε επίσης στο παραπάνω παράδειγμα πως χρειαζόμαστε και τον πίνακα αρχικών πιθανοτήτων π ο οποίος είναι ο παρακάτω:

$$\pi=[0.6\ 0.4]$$

Η διαδικασία που θα ακολουθηθεί στην συνέχεια εφόσον έχουμε περιγράψει το μοντέλο, είναι να υπολογίσουμε τις πιθανότητες να προκύψουν συγκεκριμένες ακολουθίες και συγκεκριμένα την πιο πιθανή σειρά που μας δίνει μια ακολυθία. Δηλαδή έστω ότι έχουμε την ακολουθία O={S,M,S,L} στο παράδειγμα μας, και το μοντέλο που περιγράψαμε παραπάνω. Η εύρεση της πιο πιθανής παραγωγής αυτής της ακολουθίας δεν είναι τόσο εύκολη, και μάλιστα εκτός από την χρήση των HMM ακόμα και τεχνικές δυναμικού προγραμματισμού θα μπορούσαν να χρησιμοποιηθούν (αναδρομική αναζήτηση) με τα αποτελέσματα τους να διαφέρουν σε κάποιες περιπτώσεις. Θα προχωρήσουμε στην ανάλυση του παραδείγματος που έχουμε αναφέρει και πως μπορούμε να υπολογίσουμε την πιο πιθανή περίπτωση να προκύψει αυτή η ακολουθία που αναφέραμε. Για να διευκολυθούμε αναλύουμε τα διάφορα επιμέρους στοιχεία του μοντέλου ως εξής:

T = μήκος της ακολουθίας που μας ενδιαφέρει

N = ο αριθμός των καταστάσεων του μοντέλου

M = ο αριθμός των διαφορετικών παρατηρήσεων που υπάρχουν

Q = $\{q_0,q_1,.....,q_{n-1}\}$ = ξεχωριστές καταστάσεις της διαδικασίας

V = {0,1,.....,M-1} = πιθανές παρατηρήσεις

A = πίνακας μετάβασης καταστάσεων

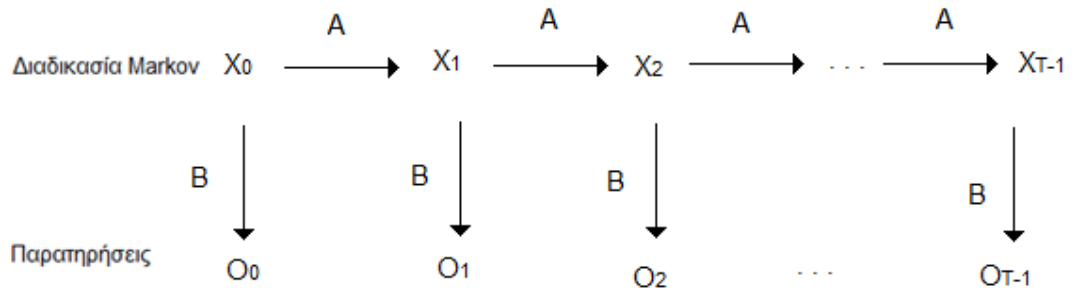
B = πίνακας πυκνότητας πιθανότητας

π = πίνακας αρχικών πιθανοτήτων

O = $(O_0,O_1,.....,O_{T-1})$ = ακολουθία παρατηρήσεων:

Όπου $O_0,O_1,....,O_{T-1}$ Ε V

Και παρακάτω δίνεται ένα γενικό HMM :



Εικόνα 5 – Αναπαράσταση HMM

Όπου Xi, είναι η ακολουθία των κρυφών καταστάσεων (Hidden states).

Έτσι για την ακολουθία που αναφέραμε πριν και αναλύουμε τώρα (την O={S,M,S,L}), θα έχουμε T = 4, N = 2, M = 3, Q = {H,C}, V = {0, 1, 2} (όπου 0,1,2 είναι οι τιμές S,M,L αντίστοιχα) και οι πίνακες A και B, όπως τους δείξαμε παραπάνω, με τον A να είναι NxN και τον B να είναι NxM. Έτσι, η πιθανότητα για να προκύψει μια ακολουθία τεσσάρων παρατηρήσεων σαν αυτή που μας ενδιαφέρει υπολογίζεται ως εξής:

$$P(X)=\pi_{x0}b_{x0}(O_0)a_{x0,x1}b_{x1}(O_1)a_{x1,x2}b_{x2}(O_2)a_{x2,x3}b_{x3}(O_3)$$

Και η πιθανότητα για την ακολουθία που μας ενδιαφέρει είναι συγκεκριμένα :

$$P(HHCC)=0.6(0.1)(0.7)(0.4)(0.3)(0.7)(0.6)(0.1)=0.000212$$

Πάνω σε αυτή τη θεωρία θα βασιστούμε και εμείς έτσι ώστε να προκύψει το μοντέλο που θέλουμε στο επόμενο κεφάλαιο και αξιολόγηση του μέσω του αλγορίθμου Forward που αναφέραμε και προηγουμένως.

Στην συγκεκριμένη όμως περίπτωση που θα αναλύσουμε, η κατάσταση έχει ως εξής. Έχουμε στην διάθεση μας δεδομένα, και θέλουμε να κατασκευάσουμε το μοντέλο που περιγράψαμε παραπάνω. Όμως δεν γνωρίζουμε και τον αριθμό των κρυφών καταστάσεων που έχουμε, έτσι και θα πρέπει να πειραματιστούμε για να δούμε ποιος είναι ο καταλληλος αριθμός κρυφών καταστάσεων στην περίπτωση μας. Και αυτό θα γίνει όπως θα δούμε παρακάτω με αξιολόγηση των παραγόμενων μοντέλων που θα έχουν διαφορετικό αριθμό κρυφών καταστάσεων. Γενικά πάντως , κάθε πρόβλημα που επιλύεται με τα HMM μπορεί να περιγραφεί ως ένα από τα παρακάτω:

1)Έχοντας το μοντέλο λ=(A,B,π) και μια ακολουθία παρατηρήσεων O θέλουμε να βρούμε την πιθανότητα P(O|λ), δηλαδή ψάχνουμε την πιθανότητα να προκύψει μια δεδομένη ακολουθία με βάση κάποιο μοντέλο λ.

2)Έχοντας πάλι το μοντέλο λ=(A,B,π) και μια ακολουθία παρατηρήσεων O θέλουμε να βρούμε την βέλτιστη ακολουθία που περιγράφει την διαδικασία Markov. Δηλαδή αναζητούμε τις κρυφές καταστάσεις του μοντέλου.

3)Έχοντας μια ακολουθία παρατηρήσεων O και έχοντας τις διαστάσεις N και M (με βάση την παραπάνω θεωρία) , θέλουμε να βρούμε το μοντέλο λ=(A,B,π) το οποίο μεγιστοποιεί την πιθανότητα να προκύψει η ακολουθία O. Δηλαδή η εκπαίδευση του μοντέλου.

Άρα από τα παραπάνω βλέπουμε πως το κομμάτι που πρέπει να ασχοληθούμε είναι συνδυασμός των παραπάνω προβλημάτων , φυσικά με σωστή σειρά, εφόσον στην αρχή του πειράματος δεν είχαμε ούτε το μοντέλο.

4.2 Χρήσιμοι αλγόριθμοι για την εφαρμογή των Hidden Markov Models

Κρίνεται απαραίτητο να δωθεί επίσης το θεωρητικό υπόβαθρο και με την χρήση παραδειγμάτων, των βασικών αλγορίθμων που χρειάζονται για την επίλυση προβλημάτων σχετικών με τα Hidden Markov Models. Οι αλγόριθμοι που θα αναλύσουμε είναι απαραίτητοι για την εκπαίδευση καθώς και για την αξιολόγηση του μοντέλου, και είναι ο Viterbi καθώς και ο Forward.

Ο αλγόριθμος Viterbi βασίζεται στην εξής θεωρία. Έστω ότι έχουμε ένα HMM με χώρο καταστάσεων S, και αρχικές πιθανότητες πi, καθώς και πίνακα μετάβασης καταστάσεων αi,j όπως ακριβώς έχουμε περιγράψει και στην θεωρία, και έχουμε παρατηρήσεις y1,…,yt. Η πιο πιθανή ακολουθία xi,…..,xt δίνεται από τα παρακάτω :

$V_{1,k} = P(y_1|k)\pi_\kappa$

$V_{t,k} = P(y_t|k)\max_{xES}(a_{x,k} * V_{t-1,x})$

Όπου $V_{t,k}$ είναι η πιθανότητα της πιο αναμενόμενης ακολουθίας για τις πρώτες t παρατηρήσεις και η οποία έχει ως τελική κατάσταση την k. Ο αλγόριθμος Viterbi υλοποιείται με την χρήση της κατάστασης x (της $2^{ης}$ εξίσωσης). Η εξίσωση η οποία μας δίνει την τιμή της x είναι Ptr(k,t), και υπολογίζουμε το $V_{t,k}$ εάν έχουμε t>1 ή αλλιώς κρατάμε το k εάν t=1, και τα ορίζουμε ως εξής.

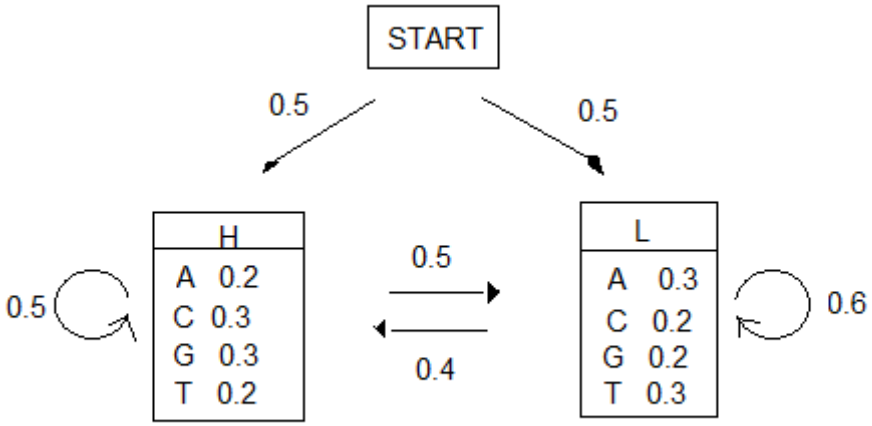
$x_T = \text{argmax}_{xES}(V_{T,x})$

$x_{t-1}=Ptr(x_t,t)$

Και τέλος να σημειώσουμε πως η πολυπλοκότητα του αλγορίθμου είναι $O(T \times |S|^2)$

Για να γίνει ευκολότερα κατανοητή η διαδικασία του αλγορίθμου δίνουμε και ένα παράδειγμα.

Έστω ότι έχουμε το παρακάτω σύστημα:



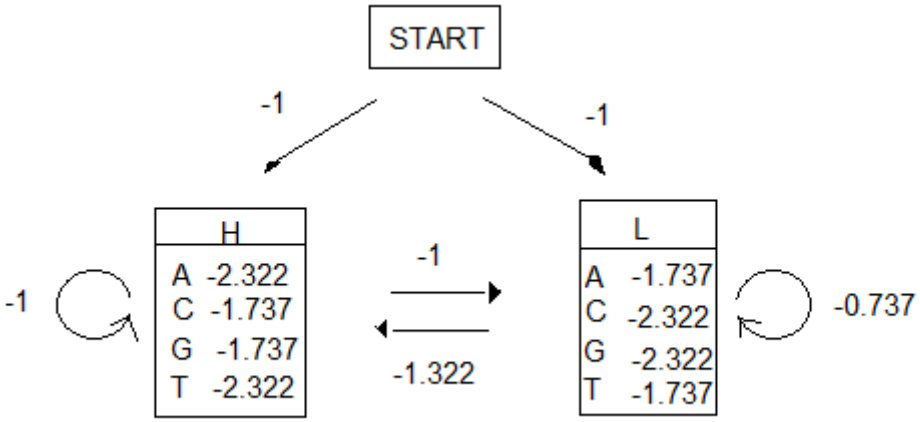
Εικόνα 6 – Πιθανότητες μεταβάσεων

Το οποίο HMM έχει δύο καταστάσεις την H και την L. Έστω ότι έχουμε την ακολουθία, S=GGCACTGAA στην οποία μπορούμε να φτάσουμε με πολλούς διαφορετικούς τρόπους μέσω των καταστάσεων L και H, όπως μέσω της P=LLHHHLLL. Στην περίπτωση αυτή η πιθανότητα να παραχθεί η ακολουθία S από το μονοπάτι P είναι :

p = pL(0) * pL(G) * pLL * pL(G) * pLH * pH(C) * ...

= 0.5 * 0.2 * 0.6 * 0.2 * 0.4 * 0.3 * ...=….

Σε αυτό το σημείο του HMM, χρειαζόμαστε τον αλγόριθμο Viterbi, ο οποίος είναι ένας αλγόριθμος βασισμένος στον δυναμικό προγραμματισμό και μας επιτρέπει να υπολογίσουμε το πιο πιθανό μονοπάτι για να συμβεί η ακολουθία αυτή. Έτσι σύμφωνα με τους παραπάνω τύπους που δώσαμε κατά το θεωρητικό κομμάτι του αλγορίθμου υπολογίζουμε τις πιθανότητες που προκύπτουν και έχουμε το εξής:



Εικόνα 7 – Πιθανότητες μεταβάσεων

Έχοντας πλέον τις πιθανότητες αυτές υπολογισμένες όπως παραπάνω ο αλγόριθμος αρχίζει να υπολογίζει τις πιθανότητες για να προκύψει η ακολουθία από το πιο πιθανό μονοπάτι.

Έτσι για την S=GGCACTGAA, αρχικά για το G , έχουμε πιθανότητα από την κατάσταση H pH(G,1) = -1 -1.737 = -2.737, και πιθανότητα από την κατάσταση L,pL(G,1) = -1 -2.322 = -3.322. Ομοίως για την ακολουθία GG, έχουμε πιθανότητες αντίστοιχα,

pH(G,2) = -1.737 + max (pH(G,1)+pHH, pL(G,1)+pLH)= -1.737 + max (-2.737 -1 , -3.322 -1.322) = -5.474 (από το pH(G,1)) και

pL(G,2) = -2.322 + max (pH(G,1)+pHL, pL(G,1)+pLL)

= -2.322 + max (-2.737 -1.322 , -3.322 -0.737)= -6.059 (από το pH(G,1))

Έτσι για όλη την ακολουθία θα προκύψει ο παρακάτω πίνακας με την τεχνική back-tracking:

	G	G	C	A	C	T	G	A	A
H	-2.73	-5.47	-8.21	-11.53	-14.01	….			-25.65
L	-3.32	-6.06	-8.79	-10.94	-14.01	….			-24.49

Και κρατώντας την μεγαλύτερη τιμή βρίσκουμε το πιθανότερο μονοπάτι, όπως και φαίνεται παρακάτω:

	G	G	C	A	C	T	G	A	A
H	-**2.73**	**-5.47**	**-8.21**	-11.53	-14.01	….			-25.65
L	-3.32	-6.06	-8.79	**-10.94**	**-14.01**	….			**-24.49**

Και είναι η HHHLLLLLL με πιθανότητα -24,49 η οποία έχει όμως προκύψει με την λογαρίθμηση των πιθανοτήτων για ευκολία των χειρισμών με τις τιμές, άρα τελικά η πιθανότητα που προκύπτει από το μοντέλο θα είναι $2^{-24.49} = 4.25*10^{-8}$.

Θα προχωρήσουμε τώρα στην ανάλυση του αλγορίθμου Forward, αρχικά περιγράφοντας το θεωρητικό κομμάτι, και στην συνέχεια θα δώσουμε το αντίστοιχο παράδειγμα που δώσαμε και στον Viterbi, για να γίνει ευκολότερα κατανοητός. Η βασικός υπολογισμός του Forward γίνεται μέσω αναδρομής, και υπολογίζει τις πιθανότητες να προκύψει μια ακολουθία συγκεκριμένα ως εξής:

$$\alpha_t(x_t) = p(x_t, y_{1:t}) = \sum_{x_{t-1}} p(x_t, x_{t-1}, y_{1:t})$$

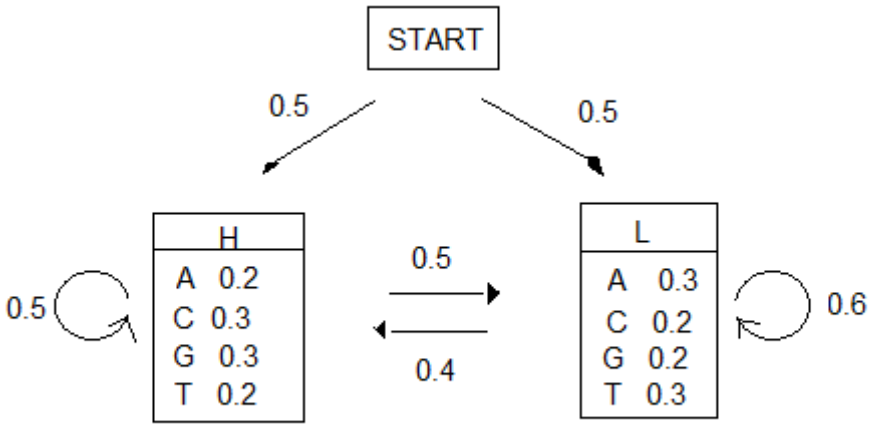
Και χρησιμοποιώντας τον κανόνα της αλυσίδας, το επεκτείνουμε ως εξής: $p(x_t, x_{t-1}, y_{1:t})$, με αποτέλεσμα ο υπολογισμός να γίνεται όπως παρακάτω:

$$\alpha_t(x_t) = \sum_{x_{t-1}} p(y_t|x_t, x_{t-1}, y_{1:t-1}) p(x_t|x_{t-1}, y_{1:t-1}) p(x_{t-1}, y_{1:t-1})$$

Και επειδή το y_t είναι ανεξάρτητο από όλες τις μεταβλητές εκτός του x_t , το οποίο x_t είναι ανεξάρτητο από όλες τις μεταβλητές εκτός του x_{t-1} η παραπάνω εξίσωση απλοποιείται στην τελική της μορφή ως εξής:

$$\alpha_t(x_t) = p(y_t|x_t) \sum_{x_{t-1}} p(x_t|x_{t-1}) \alpha_{t-1}(x_{t-1})$$

Θα χρησιμοποιήσουμε πάλι το ίδιο HMM μοντέλο που χρησιμοποιήσαμε και στην επεξήγηση του Viterbi, το οποίο θα είναι το παρακάτω:



Εικόνα 8 – Πιθανότητες μεταβάσεων

Και έστω ότι θέλουμε να υπολογίσουμε το πιθανότερο μονοπάτι για να προκύψει η ακολουθία S=GGCA . Υπολογίζουμε τις πιθανότητες οι οποίες προκύπτουν όπως στον παρακάτω πίνακα:

	START	G	G	C	A
H	0	0.5*0.3=0.15	0.15*0.5*0.3+0.1*0.4*0.3=0.0345	….+…..	0.0013767
L	0	0.5*0.2=0.1	0.1*0.6*0.2+0.15*0.5*0.2=0.027	….+…..	0.0024665

Και το άθροισμα των τελικών πιθανοτήτων θα είναι P(S)=0.0038432

Ανακεφαλαιώνοντας, από την ανάλυση αυτή προκύπτει ότι ο Viterbi είναι ο κατάλληλος αλγόριθμος βασικά για να υπολογίζουμε το πιο πιθανό μονοπάτι, καθώς και την πιθανότητα. Βέβαια για να χρησιμοποιηθεί θα πρέπει να γνωρίζουμε τις παραμέτρους του HMM μοντέλου καθώς και μια ακολουθία. Και ουσιαστικά βρίσκει την μέγιστη πιθανότητα από μια σειρά πιθανών ακολουθιών. Ο Forward από την άλλη υπολογίζει την πιθανότητα να συμβεί μια ακολουθία χρησιμοποιώντας το άθροισμα των επιμέρους ακολουθιών, και είναι πολύ χρήσιμο εργαλείο για την αξιολόγηση των παραγόμενων μοντέλων μας.

Στην συνέχεια θα εξηγήσουμε την διαδικασία του cross-validation, που μας είναι απαραίτητη για την αξιολόγηση των παραγόμενων μοντέλων, και θα δούμε τις παραλλαγές της καθώς και που μας εξυπηρετεί στην μελέτη μας.

Η τεχνική του Cross-validation, είναι ένα μοντέλο μιας τεχνικής που προσφέρει έλεγχο και αξιολογεί πως τα αποτελέσματα μιας στατιστικής ανάλυσης [20] όπως τα HMM γενικεύονται σε σχέση με ένα ανεξάρτητο κομμάτι δεδομένων. Χρησιμοποιείται ευρέως σε εφαρμογές όπου ο σκοπός τους είναι η πρόβλεψη και όπου ο ερευνητής θέλει να βρει με πόση ακρίβεια ένα μοντέλο προβλέψεων δίνει αποτελέσματα. Σε ένα πρόβλημα προβλέψεων το μοντέλο αποτελείται κυρίως από ένα σύνολο δεδομένων το οποίο πρέπει πρώτα να εκπαιδευτεί, όπως ακριβώς θα κάνουμε και εμείς με την χρήση των HMM και στην συνέχεια ένα σύνολο άγνωστων δεδομένων θα ελεγχθεί με βάση το μοντέλο που έχουμε, ώστε να βρούμε πόσο αποδοτικό είναι. Ο σκοπός είναι με την χρήση του cross-validation να μπορέσουμε να κάνουμε την διαδικασία ελέγχου κατά την φάση εκπαίδευσης του μοντέλου και να περιορίσουμε πιθανά προβλήματα ώστε το μοντέλο μας να αποδίδει σωστά σε όλα τα ανεξάρτητα δεδομένα.

Η πρώτη επανάληψη περιλαμβάνει τον διαχωρισμό των δεδομένων σε υποσύνολα, στα οποία θα εφαρμόσουμε ξεχωριστά στο καθένα στη συνέχεια ανάλυση και στα υπόλοιπα δεδομένα θα ελέγχουμε τα αποτελέσματα τους. Όπως γίνεται κατανοητό, με την χρήση μονάχα μίας επανάληψης είναι πολύ πιθανόν τα αποτελέσματα μας να μην είναι ακριβή για όλο το δείγμα, έτσι για να μειώσουμε αυτή την μεταβλητότητα, είμαστε αναγκασμένοι να προβούμε σε πολλές επαναλήψεις cross-validation σε διάφορα υποσύνολα, και να πάρουμε τα τελικά αποτελέσματα, από τον μέσο όρο των επαναλήψεων αυτών. Γίνεται εμφανές επίσης, ότι η χρήση του cross-validation είναι πολύ σημαντική τεχνική ώστε να

αξιολογούμε σωστά τα μοντέλα, και να αποφεύγουμε λάθη σε δείγματα, ή να αποφεύγουμε χρονοβόρες διαδικασίες μελέτης μεγάλων δειγμάτων.

Επίσης υπάρχουν διάφορα είδη τεχνικών cross-validation τα οποία χωρίζονται σε δύο κατηγορίες. Στις εξαντλητικές τεχνικές όπου μαθαίνουμε και ελέγχουμε το μοντέλο με όλους τους δυνατούς τρόπους που μπορούμε να χωρίσουμε σε υποσύνολα το αρχικό δείγμα στο κομμάτι της εκπαίδευσης και στο υποσύνολο ελέγχου. Και στις μη εξαντλητικές τεχνικές όπου δεν υπολογίζουμε σε κάθε υποσύνολο του αρχικού δείγματος. Στις εξανλητικές μεθόδους έχουμε τις εξής επιλογές , να αφήνουμε p υποσύνολα εκτός (leave-p-out cross validation) , η οποία τεχνική αφορά p παρατηρήσεις ως το υποσύνολο, και τις υπόλοιπες τις χρησιμοποιούμε για εκπαίδευση του μοντέλου. Και η διαδικασία αυτή επαναλαμβάνεται με όλους τους πιθανούς τρόπους να χωρίσουμε το αρχικό δείγμα. Η άλλη τεχνική που έχουμε είναι να αφήνουμε ένα υποσύνολο εκτός κάθε φορά(leave-one-out cross validation) , όπου έχουμε ουσιαστικά υποπερίπτωση της παραπάνω με p=1, και αφήνουμε κάθε φορά εκτός μονάχα ένα υποσύνολο, για όλα τα πιθανά υποσύνολα(η μέθοδος αυτή θα χρησιμοποιηθεί και για αξιολόγηση στα πειράματα μας)

Στις μη εξανλτητικές μεθόδους, έχουμε τώρα την γενική περίπτωση των k αναδιπλώσεων (k-fold cross validation) όπου το αρχικό δείγμα χωρίζεται τυχαία σε k υποσύνολα ίδιου μεγέθους. Από αυτά τα υποσύνολα, ένα υποσύνολο το κρατάμε για αξιολόγηση και τα υπόλοιπα k-1 υποσύνολα τα χρησιμοποιούμε για την εκπαίδευση του μοντέλου, και η διαδικασία αυτή συνεχίζεται για k φορές, όσα είναι δηλαδή και τα υποσύνολα. Στην συνέχεια από τα k αποτελέσματα από τις αναδιπλώσεις υπολογίζουμε την μέση τιμή για να βρούμε μια προσέγγιση. Το πλεονέκτημα αυτής της μεθόδου , σε σχέση με την επαναλαμβανόμενη τυχαία χρήση των υποσυνόλων είναι ότι όλες οι παρατηρήσεις χρησιμοποιούνται και για εκπαίδευση αλλά και για έλεγχο, και κάθε παρατήρηση χρησιμοποιείται για έλεγχο ακριβώς μία φορα. Μια πολύ συνηθισμένη χρήση της τεχνικής αυτής είναι με η τεχνική με 10 αναδιπλώσεις , όμως η παράμετρος k εξαρτάται από τον εκάστοτε ερευνητή. Η συγκεκριμένη μέθοδος είναι και αυτή που θα χρησιμοποιήσουμε σε ένα τμήμα των πειραμάτων μας.

Επίσης όταν ο αριθμός των αναδιπλώσεων ισούται με τον αριθμό των παρατηρήσεων , δηλαδή k=n τότε η μέθοδος αυτή εμπίπτει στην περίπτωση της εξαντλητικής μεθόδου που αφήνει ένα υποσύνολο εκτός κάθε φορά.

Τέλος, υπάρχουν και άλλες τεχνικές μη εξαντλητικές όπως με δύο αναδιπλώσεις, που είναι υποπερίπτωση της παραπάνω, καθώς και επαναλαμβανόμενος έλεγχος με την χρήση τυχαίων υποσυνόλων όπου χωρίζεται το δείγμα τυχαία, σε υποσύνολα άλλα για εκπαίδευση και άλλα για αξιολόγηση.

4.3 Τα βήματα που θα ακολουθήσουμε

Έχοντας πλέον δώσει το απαραίτητο θεωρητικό υπόβαθρο, και έχοντας αναφέρει όσα χρειάζονται για την προσέγγιση μας σε αυτή την εργασία, θα εξηγήσουμε πως θα συνδυάσουμε τα βήματα αυτά,σε αυτό το κομμάτι για να γίνουν ευκολότερα κατανοητά.

Αρχικά βασιζόμενοι στο μοντέλο CRUISE θα πάρουμε τα δεδομένα από τις συνεδρίες των χρηστών, οι οποίες και θα αποθηκευτούν στην βάση δεδομένων μας. Στην συνέχεια θα επεξεργαστούμε αυτά τα δεδομένα και θα κατασκευάσουμε διάφορα HMM από τα οποία και θα επιλέξουμε το καταλληλότερο για την εφαρμογή μας, μέσα από την διαδικασία της αξιολόγησης.

Στο επόμενο κεφάλαιο, θα δούμε στην πράξη πως θα γίνει η υλοποίηση της εργασίας, δίνοντας όλες τις απαραίτητες τεχνικές πληροφορίες που χρειάστηκαν για να το κατασκευάσουμε. Τα τεχνικά αυτά θέματα αφορούν, ανάλυση της μηχανής CRUISE πάνω στην οποία βασιστήκαμε, και επίσης τι χρειαστήκαμε για την εξόρρυξη των πληροφοριών. Επίσης, ανάλυση της βάσης δεδομένων μας , καθώς και όλη την υλοποίηση των μοντέλων με την βοήθεια του Matlab.

Κεφάλαιο 5

Υλοποίηση με το CRUISE και με χρήση των HMM

5.1 Η υλοποίηση μας μέσω του CRUISE

Για την υλοποίηση και την αξιοποίηση στην συνέχεια υλικού από τις συνήθειες των χρηστών σε μια συγκεκριμένη εφαρμογή, όπως αναφέραμε και στο εισαγωγικό κομμάτι θα χρησιμοποιήσουμε την εφαρμογή CRUISE η οποία μας δίνει την δυνατότητα, να κάνουμε αναζήτηση σε διάφορους τομείς όπως σε tweets του μέσου κοινωνικής δικτύωσης Twitter, ή αναζητήσεις εικόνων και βίντεο από την ειδική ιστοσελίδα της Yahoo, το Flickr, καθώς και συνδέσμων(links) για θέματα που μας ενδιαφέρουν. Επίσης στο CRUISE έχουμε την δυνατότητα να φιλτράρουμε το περιβάλλον από το οποίο θα συλλέγουμε το υλικό αυτό , σαν φιλτράρισμα των αναζητήσεων μας, πχ με το αν οι ειδήσεις θα προέρχονται από τον δικο μας λογαριασμό Twitter, είτε από φίλους μας, ή από ολόκληρη την ιστοσελίδα. Επίσης ο χρήστης έχει την δυνατότητα να κάνει διερευνητική αναζήτηση μέσω του CRUISE εφόσον υπάρχει η δυνατότητα , μέσα από κάθε αναζήτηση του να κρατάει τα δεδομένα τα οποία θεωρεί χρήσιμα στην αναζήτηση του, και σε κάθε επόμενη αναζήτηση, να γίνεται φιλτράρισμα από αυτά και όσο προχωράει η αναζήτηση να γίνεται όλο και πιο εξειδικευμένη.

Συγκεκριμένα, κατά την έναρξη του CRUISE, ο χρήστης έχει την δυνατότητα να ξεκινήσει μια αναζήτηση, και να επιλέξει από που θα αντλήσει υλικό για την αναζήτηση του, από τέσσερις διαφορετικές επιλογές. Αυτές μπορεί να είναι το Twitter, εικόνες, σύνδεσμοι, ή βίντεο, και φυσικά όχι μόνο ένα από αυτά αλλά και συνδυασμός τους. Αφού έχει γράψει κάποιον όρο ή όρους αναζήτησης και επιλέξει την αναζήτηση (search) , θα εμφανιστούν ακριβώς από κάτω τα αποτελέσματα. Τα αποτελέσματα ποικίλουν φυσικά ανάλογα τις πηγές που έχει επιλέξει. Έχοντας κάνει πλέον την αρχική αναζήτηση, θα αρχίσει ουσιαστικά το κομμάτι της διευρευνητικής αναζήτησης, όπου άλλα στοιχεία θα επιλέξει να κρατήσει και άλλα όχι. Συγκεκριμένα, από τα αποτελέσματα που έχουν εμφανιστεί μπορεί να διαλέξει ποια θέλει

να κρατήσει κατά την συνέχεια της αναζήτησης του.

Αν τον ενδιαφέρει ένα αποτέλεσμα που είναι από το Twitter έχει την δυνατότητα περνώντας τον κέρσορα από πάνω του(hover) να επιλέξει το κουμπί search με το οποίο θα του εμφανιστεί ολόκληρο το tweet ώστε να κρίνει αν όντως το αποτέλεσμα που προέκυψε είναι σχετικό με αυτό που ψάχνει.

Εάν τώρα, το αποτέλεσμα είναι οποιασδήποτε από τις παραπάνω κατηγορίες έχει δύο δυνατότητες περνώντας τον κέρσορα πάνω από το αποτέλεσμα. Αυτές είναι, αφενός η προσθήκη του συγκεκριμένου αποτελέσματος στην λίστα του, καθώς και το άνοιγμα του όρου σε νέο παράθυρο και συγκεκριμένα άνοιγμα της πηγής από όπου βρήκαμε το συγκεκριμένο αποτέλεσμα.

Έτσι, προχωράμε την αναζήτηση μας με το CRUISE με αυτόν τον τρόπο, προσθέτοντας αποτελέσματα στην λίστα μας, και επιλέγοντας συνεχώς νέες αναζητήσεις όλο και πιο εξειδικευμένες μέχρι να πάρουμε τα επιθυμητά αποτελέσματα.

5.2 Αναλυτική παρουσιάση των κινήσεων του χρήστη

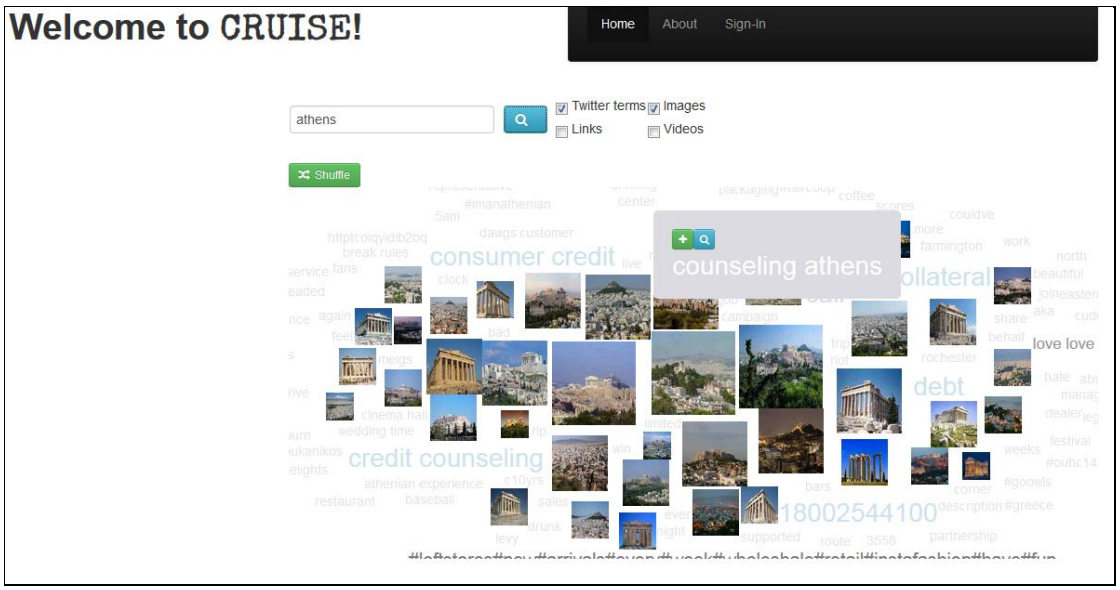
Στην συνέχεια θα αναλύσουμε επακριβώς τα σημεία ενδιαφέροντος με τα οποία ασχοληθήκαμε στην συγκεκριμένη εργασία όσον αφορά τις κινήσεις των χρηστών στο CRUISE.

Οι συνολικές κινήσεις των χρηστών που μας ενδιαφέρουν και αποθηκεύουμε , αποτελούνται από 15 διαφορετικές κινήσεις, οι οποίες και αναλύονται και κωδικοποιούνται στα παρακάτω:

Κίνηση(Action)	Αριθμός
New Query	1
Click on Term for Info	2
Click on Image for Info	3
Click on Video for Info	4
Click on Link for Info	5
Click on Image to View	6
Click on Video to Watch	7
Click on Link to Read	8
Add term to Bucket	9
Add image to Bucket	10
Add video to Bucket	11
Add link to Bucket	12
Idle User	13
Start Exploration	14
Shuffle	15

Πίνακας 1 - Πίνακας κινήσεων των χρηστών

Παρακάτω δίνεται ένα δείγμα αναζήτησης του CRUISE, και για διευκόλυνση θα εξηγήσουμε τι ακριβώς είναι η κάθε επιλογή στην εφαρμογή ώστε να γίνει ευκολότερα κατανοητό.



Εικόνα 9- Στιγμιότυπο της εφαρμογής CRUISE

Ξεκινόντας από πάνω αριστερά, έχουμε τον χώρο που γράφουμε τους όρους που θέλουμε να κάνουμε αναζήτηση και ακριβώς στα δεξιά του έχουμε το κουμπί της αναζήτησης με το οποίο και γίνεται η κίνηση 1 , New Query(Νέα αναζήτηση). Δεξιά από το κουμπί της αναζήτησης έχουμε τέσσερις επιλογές που είναι να επιλέξουμε τι είδους αποτελέσματα

θέλουμε να εμφανιστούν(επιλογή ανάμεσα σε twitter,εικόνες,σύνδεσμοι,βίντεο), και στην συγκεκριμένη λήψη έχουμε επιλέξει να πάρουμε όρους από το twitter, καθώς και εικόνες. Επίσης, με πράσινο φόντο κάτω από την μπάρα αναζήτησης έχουμε το κουμπί Shuffle που ανακατανέμει τα αποτελέσματα. Ακριβώς από κάτω από αυτά έχουμε την περιοχή των αποτελεσμάτων, που όπως βλέπουμε αποτελείται από όρους του twitter και εικόνες. Αν τοποθετήσουμε τώρα τον κέρσορα πάνω από κάποιον όρο, τότε εμφανίζονται οι επιλογές της προσθήκης στη συλλογή μου (+) καθώς και των περισσοτέρων πληροφοριών δίπλα του. Κάτω από τα αποτελέσματα, έχουμε τα κουμπιά Share και Explore. Το Share, μοιράζει τα στοιχεία που έχουμε προσθέσει στη συλλογή μας, σε άλλες ιστοσελίδες, ή μέσα κοινωνικής δικτύωσης, και το Explore είναι η κίνηση για την έναρξη νέας αναζήτησης με τους όρους που έχουμε ήδη προσθέσει στην συλλογή μας. Τέλος, κάτω από αυτά τα κουμπιά έχουμε τους όρους που έχουμε προσθέσει από τα αποτελέσματα. Φυσικά στην κορυφή της σελίδας υπάρχουν και οι επιλογές για πληροφορίες για την εφαρμογή καθώς και εγγραφή του χρήστη.

Στην συνέχεια, αφού έχουμε δώσει μια γενική εικόνα για την εφαρμογή πάνω στην οποία βασιστήκαμε, θα αναλύσουμε κάθε μία από τις παραπάνω κινήσεις επακριβώς, πως προκύπτουν, σε τι εξυπηρετούν στο CRUISE καθώς και ποιες πληροφορίες κρατάμε στην βάση δεδομένων για αξιοποίηση τους στην συνέχεια που θα επεξεργαστούμε τα δεδομένα αυτά.

Νέα αναζήτηση-New Query

Είναι η νέα αναζήτηση του χρήστη(καθώς και η αρχική). Κατά την νέα αναζήτηση οι επιλογές του χρήστη- πληροφορίες που κρατάμε είναι οι όροι της αναζήτησης (terms) , καθώς και η πηγή-πηγές από όπου προέρχονται αυτές οι πληροφορίες. Οι όροι μπορεί να είναι ένας ή περισσότεροι, όπως και οι πηγές από όπου θα προέρχονται τα αποτελέσματα θα είναι το twitter, εικόνες, βίντεο , σύνδεσμοι ή συνδυασμός αυτών. Μαζί με αυτά τα στοιχεία θα κρατάμε και την ακριβή ημερομηνία και ώρα που έγινε η αναζήτηση.

Πχ μια πιθανή εγγραφή θα είναι :

Id | New Query | terms | source | timestamp

"2873";"New Query|twitter,images|weather,athens |Mon Jul 28 2014 21:05:20 GMT+0300 (GTB Standard Time)"

Κλικ σε Όρο για πληροφορίες/Κλικ σε Εικόνα για πληροφορίες/Κλικ σε Σύνδεσμο για πληροφορίες/Κλικ σε Βίντεο για πληροφορίες-

Click on Term for Info/Click on Image for Info/Click on Link for Info/Click on Video for Info

Κατά το πάτημα του κέρσορα του ποντικιού του χρήστη στο ειδικό κουμπί info πάνω σε έναν όρο ή σε μια εικόνα,βίντεο ή σύνδεσμο (αφού έχει πρώτα τοποθετήσει τον κέρσορα πάνω από αυτό για να εμφανιστεί η επιλογή πάτημα για πληροφορίες) ο χρήστης έχει την δυνατότητα να δει περισσότερες πληροφορίες σχετικά με όποιο αποτέλεσμα επιθυμεί. Έτσι κατά την κάθε συγκεκριμένη κίνηση από τον χρήστη εμείς θα κρατάμε τις παραπάνω πληροφορίες καθώς και την στιγμή που έγινε αυτή όπως φαίνεται παρακάτω.

Πχ μια πιθανή εγγραφή θα είναι :

Id | Click for info | source | terms | timestamp

"2874";"Click for info|Text|movies|Mon Jul 28 2014 21:05:29 GMT+0300 (GTB Standard Time)"

Κλικ σε Εικόνα για προβολή/Κλικ σε Βίντεο για προβολή/Κλικ σε Σύνδεσμο για ανάγνωση- Click on Image to View/Click on Video to Watch/Click on Link to Read

Εάν ο χρήστης πάει τον κέρσορα πάνω από μια εικόνα, βίντεο ή σύνδεσμο, τότε έχει την δυνατότητα να κάνει κλικ ακριβώς πάνω σε αυτό , σε οποιδήποτε σημείο του με αποτέλεσμα να ανοίξει το αντίστοιχο αποτέλεσμα σε νέα καρτέλα του φυλλομετρητή που χρησιμοποιούμε, και συγκεκριμένα να ανοίξει η πηγή από την οποία βρήκαμε το αποτέλεσμα αυτό. Και εμείς θα κρατάμε τις παραπάνω πληροφορίες καθώς και την στιγμή που έγινε αυτή η επιλογή.

Πχ μια πιθανή εγγραφή θα είναι :

Id | Click on | source | terms | timestamp

"2875";"Click on|Image|movies james bond |Mon Jul 28 2014 21:05:32 GMT+0300 (GTB Standard Time)"

Προσθήκη όρου στην λίστα/Προσθήκη Εικόνας στην λίστα/Προσθήκη Βίντεο στην λίστα/Προσθήκη Συνδέσμου στην λίστα-

Add term to Bucket/Add image to Bucket/Add video to Bucket/Add link to Bucket

Ομοίως αφού ο χρήστης έχει τοποθετήσει τον κέρσορα πάνω από κάποιο αποτέλεσμα ώστε να εμφανιστεί πρώτα το αντίστοιχο σύμβολο για επιλογή προσθήκη σε, έχει την δυνατότητα πατώντας το να προσθέσει το συγκεκριμένο αποτέλεσμα στην λίστα του ώστε κατά την νέα αναζήτηση του να το συμπεριλάβει και αυτό. Προφανώς προϋπόθεση για να είναι χρήσιμη αυτή η επιλογή είναι η παρακάτω δυνατότητα που θα αναλύσουμε, που δεν είναι άλλη από την Start Exploration, δηλαδή την νέα αναζήτηση από τους υπάρχοντες όρους στην λίστα μας.

Κατά την προσθήκη του όρου στη λίστα μας, κρατούμε όπως και στις προηγούμενες περιπτώσεις με το Click όλες αυτές τις πληροφορίες που περιγράψαμε καθώς και τον χρόνο που συνέβη.

Πχ μια πιθανή εγγραφή θα είναι :

Id | Add to bucket | source | terms | timestamp

"2879";"Add to bucket|Image|weather athens |Mon Jul 28 2014 21:15:33 GMT+0300 (GTB Standard Time)"

Αδράνεια - Idle User

Η κίνηση αυτή συμβαίνει όταν ο χρήστης παραμένει ανενεργός για κάποιο χρονικό διάστημα και συγκεκριμένα έχουμε ορίσει για 20 δευτερόλεπτα. Δηλαδή δεν υπάρχει καμία κίνηση ούτε του κέρσορα ούτε κάποια αναζήτηση. Η χρησιμότητα αυτής της κίνησης εξυπηρετεί κυρίως για να αξιολογήσουμε την εμπειρία χρήσης του CRUISE και πότε ο χρήστης δεν βρίσκεται στην σελίδα. Επίσης και εδώ κρατάμε την χρονική στιγμή που συνέβη αυτή η κίνηση.

Πχ μια πιθανή εγγραφή θα είναι :

Id | Idle user | timestamp

"2881";"Idle user |Mon Jul 28 2014 21:15:33 GMT+0300 (GTB Standard Time)"

Έναρξη αναζήτησης - Start Exploration

Η έναρξη αναζήτησης συμβαίνει όταν ο χρήστης έχει ήδη προσθέσει κάποιους όρους στη λίστα αυτών που τον ενδιαφέρουν, και πραγματοποιείται νέα αναζήτηση σε νέα καρτέλα που περιλαμβάνει τους όρους αυτούς. Κατά την κίνηση αυτή επίσης κρατάμε τα δεδομένα που έχει προσθέσει ο χρήστης στη λίστα του, καθώς και την χρονική στιγμή που συνέβη.

Πχ μια πιθανή εγγραφή θα είναι :

Id | Start Exploration | source | terms | timestamp

"2982";" Start Exploration |Image|average temperature in Greece |Mon Jul 28 2014 21:17:33 GMT+0300 (GTB Standard Time)"

Ανακάτεμα - Shuffle

Η κίνηση Shuffle είναι το ανακάτεμα των αναζητήσεων και εξυπηρετεί τον χρήστη σε περίπτωση που θέλει να του παρουσιαστούν τα αποτελέσματα με άλλο τρόπο στο κέντρο της οθόνης. Κρατάμε την κίνηση καθώς και την χρονική στιγμή.

Πχ μια πιθανή εγγραφή θα είναι :

Id | Shuffle | timestamp

"2882";" Shuffle |Mon Jul 28 2014 21:25:33 GMT+0300 (GTB Standard Time)"

Όλες οι παραπάνω κινήσεις έγιναν με την χρήση Javascript και συγκεκριμένα της βιβλιοθήκης JQuery ώστε να μπορούμε να αποθηκεύουμε όλες αυτές τις πληροφορίες. Κώδικας για το πως έγιναν τα παραπάνω δίνεται στο παράρτημα.

Φυσικά πέρα από αυτές τις κινήσεις των χρηστών που τελικά αποθηκεύουμε για επεξεργασία, υπάρχουν και άλλες κινήσεις τις οποίες υλοποιήσαμε ώστε να ανακαλύψουμε συνήθειες του χρήστη που βασίζονται ακόμα και στην διάθεση του, ή εάν χρησιμοποιεί την εφαρμογή αποδοτικά.

Αρχικά είχαμε υλοποιήσει και κρατούσαμε δεδομένα για το πάνω από ποια αποτελέσματα περνάει τον κέρσορα (hover) , αλλά στην συνέχεια φάνηκε πως ενώ κάτι τέτοιο είναι αρκετά χρήσιμο για τον τρόπο χρήσης της μηχανής , ουσιαστικά μας γεμίζει με πολύ πληροφορία η οποία δεν θα είναι άμεσα χρήσιμη για επεξεργασία, καθώς εμπεριέχει πολύ θόρυβο. Ο λόγος είναι λόγω της δομής εμφάνισης των αποτελεσμάτων, ο χρήστης για να επιλέξει το αποτέλεσμα που τον ενδιαφέρει, θα πρέπει να περάσει τον κέρσορα του ακόμα και πάνω από αποτελέσματα που δεν τον ενδιαφέρουν, αλλά απλά βρίσκονταν κοντά στο επιθυμητό του αποτέλεσμα.

Επίσης, έγινε υλοποιήση και για μέτρηση της ταχύτητας του κέρσορα, ώστε να εξαγάγουμε συμπεράσματα όσον αφορά τον τρόπο χρήσης της εφαρμογής από τον χρήστη, καθώς απότομες κινήσεις ίσως φανερώνουν αδιαφορία για την εφαρμογή και όχι σωστή χρήση της. Όμως παρομοίως κάτι τέτοιο δεν μας ήταν άμεσα χρήσιμο για το επόμενο βήμα οπότε και παραλήφθηκε.

5.3 Αποθήκευση δεδομένων για επεξεργασία

Εφόσον, έχουμε πλέον αποφασίσει για την τελική μορφή των κινήσεων των χρηστών που θα κρατήσουμε και το έχουμε επίσης υλοποιήσει , πρέπει να αποθηκεύουμε όλα αυτά τα αποτελέσματα σε μια βάση δεδομένων ώστε να μπορέσουμε στην συνέχεια να τα αξιοποιήσουμε για το επόμενο βήμα. Το κομμάτι της σωστής αποθήκευσης είναι πολύ σημαντικό καθώς από εκεί εξαρτάται κυρίως το πόσο εύκολα θα μπορέσουμε στην συνέχεια να τα αξιοποιήσουμε, καθώς ο αριθμός των δεδομένων θα είναι πολύ μεγάλος και η μη αυτόματη επεξεργασία ανεύφικτη.

Για βάση δεδομένων, χρησιμοποιήσαμε την MySql η οποία και παρέχεται ως δωρεάν λογισμικό, αλλά και υπάρχει μεγάλη βοήθεια από υλικό στο διαδίκτυο. Η σύνδεση της MySql με την εφαρμογή μας γίνεται με την χρήση JPA (Java Persistence API) που δεν χρησιμοποιεί τις συμβατικές εντολές SQL όπως Insert ,Select κτλ αλλά εκμεταλλεύεται τις υπάρχουσες βιβλιοθήκες και υλοποιεί τις διάφορες κινήσεις στην βάση δεδομένων αυτόματα με χρήση κώδικα στο κομμάτι της εφαρμογής μας. Παραδείγματα εγγραφής στη βάση θα δωθούν επίσης στο παράρτημα, με τις εγγραφές που χρησιμοποιήσαμε στην εργασία αυτή.

Επίσης, απαραίτητη ήταν η αποθήκευση της χρονικής στιγμής που γίνεται η κάθε κίνηση ώστε να γνωρίζουμε την λογική ακολουθία τους, και έτσι ώστε ο αύξων αριθμός των κινήσεων στη βάση να έχει νόημα, και βλέποντας και μόνο τους αριθμούς να γνωρίζουμε με ποια σειρά έγιναν. Ένα πρόβλημα που αντιμετωπίσαμε αρχικά ήταν πως ο κάθε χρήστης πιθανόν να έχει διαφορετική ώρα , από χώρα σε χώρα, και κάτι τέτοιο προκαλούσε πρόβλημα, καθώς χανόταν η λογική ακολουθία. Έτσι, πάψαμε να χρησιμοποιούμε τοπική ώρα αλλά κοινή ώρα για όλους και συγκεκριμένα GMT , ώστε να μην υπάρχει αυτό το πρόβλημα.

Ένα ακόμα πρόβλημα το οποίο αντιμετωπίσαμε κατά την διαδικασία της εγγραφής στην βάση δεδομένων ήταν το πρόβλημα της αδυναμίας του συστήματος να γράφει ταυτόχρονα πολλαπλές εγγραφές στη βάση. Αυτό είναι ένα πολύ σημαντικό σημείο που ήθελε ιδιαίτερη προσοχή καθώς κατά την χρήση της εφαρμογής από πολλούς χρήστες ταυτόχρονα θα προκαλούσε αποτυχία υλοποίησης όλων των εγγραφών. Για να το αποφύγουμε λοιπόν, προχωρήσαμε στην λύση της μαζικής εγγραφής δεδομένων αφού συμπληρωνόταν ένας αριθμός κινήσεων. Ο αριθμός που επιλέξαμε ήταν να γίνεται η εγγραφή στη βάση δεδομένων ανά 10 κινήσεις των χρηστών, καθώς δεν θα έχουμε απώλεια πληροφορίας ανά πάσα στιγμή και επίσης το σύστημα μας μπορεί να γράφει επιτυχώς αυτό τον αριθμό αναζητήσεων ταυτόχρονα.

Στην εφαρμογή μας, ο κάθε χρήστης είναι επίσης μοναδικός , και είναι ήδη εγγεγραμμένος έχοντας το δικό του χαρακτηριστικό όνομα. Έτσι, υπάρχει η ανάγκη να αποθηκεύουμε και το όνομα του χρήστη κατά την διάρκεια μιας συνεδρίας του. Δηλαδή, κατά την αποθήκευση των δεδομένων στην βάση πέρα από τις κινήσεις του χρήστη, μας ενδιαφέρει να τις κατηγοριοποιήσουμε. Και η κατηγοριοποίηση αυτή αποτελέι να χωρίζονται οι κινήσεις ανά χρήστη, αλλά και ανά συνεδρία. Συγκεκριμένα, για να το επιτύχουμε αυτό, κατά την έναρξη μιας συνεδρίας, δίνουμε έναν αριθμό συνεδρίας, τον οποίο και κρατάμε, καθώς και το όνομα του χρήστη. Δηλαδή πέρα από τα δεδομένα που κρατάμε η βάση δεδομένων θα έχει και τα παρακάτω στοιχεία(αναλυτικότερο δείγμα δίνεται στο παράρτημα) :

163	kostis	323	1	Tue Aug 26 2014 10:13:17 GMT+0300 (GTB Standard Time)
164	kostis	323	3	Tue Aug 26 2014 10:13:32 GMT+0300 (GTB Standard Time)

Δηλαδή αύξων αριθμό της καταχώρησης, όνομα του χρήστη, μοναδικό κωδικό συνεδρίας, κίνησης του χρήστη, καθώς και ώρα. Αυτή είναι η μορφή κιόλας που μας ενδιαφέρει για την περαιτέρω επεξεργασία των δεδομένων, όπως και θα δούμε στο επόμενο κεφάλαιο.

Φυσικά, εκτός από αυτά τα δεδομένα, κρατάμε και όλα τα υπόλοιπα δεδομένα που έχουμε πάρει από την εφαρμογή όπως είδαμε σε προηγούμενη παράγραφο , δηλαδή ποιους όρους αναζήτησε ο χρήστης, τι είδους όροι ήταν κτλ. Όμως για το επόμενο βήμα, στο οποίο μας ενδιαφέρει να πάρουμε τις κινήσεις μονάχα, ανά συνεδρία χρειαζόμαστε τα δεδομένα, στην μορφή που δώσαμε παραπάνω, αφού τα έχουμε περάσει σε ένα υπολογιστικό φύλλο Excel, ώστε να διαβάσουμε τα δεδομένα στην συνέχεια από το Matlab.

5.4 Αξιοποίηση συνεδριών των χρηστών (Log)

Εφόσον, πλέον έχουμε αποθηκευμένες στη βάση δεδομένων όλες τις κινήσεις που μας ενδιαφέρουν και στην μορφή που θέλουμε για περαιτέρω χρήση τους, μπορούμε να προχωρήσουμε στο επόμενο βήμα που δεν είναι άλλο από την αξιοποίηση των δεδομένων αυτών. Τα δεδομένα αυτά περιλαμβάνουν όλες τις κινήσεις που έχουν κάνει οι χρήστες και αφορούν όλα όσα έχουμε περιγράψει στην προηγούμενη ενότητα.

Το σκεπτικό είναι να μπορέσουμε να μοντελοποιήσουμε όλα τα δεδομένα που έχουμε, ώστε να μπορέσουμε να υπολογίσουμε τις πιθανότητες να κάνει ο χρήστης την κάθε επόμενη κίνηση. Δηλαδή, σύμφωνα με την προηγούμενη συμπεριφορά των χρηστών που την έχουμε αποθηκευμένη στη βάση, θέλουμε να τροφοδοτήσουμε ένα μοντέλο με αυτά τα αποτελέσματα ώστε να μας υπολογίσει ποιες είναι οι πιθανότητες για κάθε νέα κίνηση. Καταλήξαμε πως για να το επιτύχουμε αυτό και να επεξεργαστούμε αυτά που θέλουμε μας εξυπηρετεί η χρήση των Hidden Markov Models , την θεωρία των οποίων αναλύσαμε στο προηγούμενο κεφάλαιο.

Σε αυτό το κομμάτι, θέλουμε δηλαδή να δούμε αν προκύπτει κάποια αλληλουχία ανάμεσα στις κινήσεις των χρηστών. Αν δηλαδή, ακολουθούν οι κινήσεις τους κάποιο συγκεκριμένο μοτίβο , το οποίο θα μας βοηθήσει να γνωρίζουμε πιθανές επόμενες κινήσεις τους, και κάτι τέτοιο θα ήταν πολύ χρήσιμο για βελτιώσουμε την εμπειρία χρήσης της εφαρμογής.

Αυτό που χρειαζόμαστε, για να είμαστε σε θέση να αξιοποιήσουμε τα δεδομένα είναι μονάχα τις κινήσεις των χρηστών και φυσικά να είναι ομαδοποιημένες ανά συνεδρία του χρήστη. Χρησιμοποιούμε στην ουσία δύο στήλες οι οποίες έχουν αυτά τα δύο στοιχεία, τα οποία και θα επεξεργαστούμε στην συνέχεια με την βοήθεια του Matlab. Η πρώτη στήλη με τις συνεδρίες απλά χρειάζεται για να μας βοηθήσει να κατηγοριοποιήσουμε τα χρήσιμα δεδομένα των κινήσεων των χρηστών. Η βασική στήλη είναι αυτή με τις πραγματικές κινήσεις (actions) των χρηστών από τις οποίες και θα προκύψει το τελικό μας αποτέλεσμα και αποτελείται από τις 15 πιθανές κινήσεις των χρηστών.

5.5 Επεξεργασία δεδομένων με το Matlab

Μετά την χρήση της εφαρμογής και της αποθήκευσης των δεδομένων με τον τρόπο που περιγράψαμε στο κεφάλαιο αυτό, θα έχουμε πλέον επιτυχώς αποθηκευμένες στην βάση

δεδομένων τις συνεδρίες χρηστών (user sessions) ώστε να μοντελοποιήσουμε και να εκπαιδεύσουμε το μοντέλο μας σύμφωνα με την θεωρία των HMM.

Όπως αναφέρεται και στην θεωρία των HMM, στο προηγούμενο κεφάλαιο, για να προκύψει σωστή εκπαίδευση του μοντέλου χρειάζεται πειραματισμός με διάφορες καταστάσεις και συγκεκριμένα με διαφόρους αριθμούς κρυφών καταστάσεων (Hidden States), οι οποίες και θα ποικίλουν από 2 καταστάσεις μέχρι και 7, ώστε να καλύπτουμε όλο το φάσμα με βάση τις 15 πιθανές κινήσεις των χρηστών που αντιπροσωπεύουν την εφαρμογή μας. Μετά την μοντελοποίηση θα ακολουθήσει η αναγνώριση και η εκπαίδευση η οποία και θα γίνει με την χρήση του αλγορίθμου Forward με την τεχνική του leave-one-out Cross-Validation με 231 επαναλήψεις (folds), όσες δηλαδή και οι συνεδρίες.

Αρχικά εξαγάγαμε όλα τα δεδομένα από την βάση δεδομένων στο υπολογιστικό φύλλο Excel από το οποίο και θα διαβάσει το υπολογιστικό πρόγραμμα Matlab τα αρχεία. Όπως αναφέραμε και προηγουμένως, τα ωφέλιμα δεδομένα για την εκπαίδευση του μοντέλου μας είναι οι κινήσεις του χρήστη και το πως αυτές χωρίζονται ανά χρήστη σε διάφορες συνεδρίες. Ένα πρόβλημα που αντιμετωπίσαμε αρχικά ήταν το ότι οι συνεδρίες ήταν διαφορετικού μήκους, όπως και ήταν λογικό , αφού ο κάθε χρήστης χρησιμοποιούσε την εφαρμογή για διαφορετικό χρόνο, με διαφορετικό τρόπο, άρα και σίγουρα με διαφορετικό συνολικό αριθμό κινήσεων ανά συνεδρία. Το πρόβλημα που προέκυπτε ήταν πως για τον υπολογισμό του μοντέλου χρειαζόμασταν ίδιο αριθμό κινήσεων ανά συνεδρία, όμως το να αφαιρέσουμε κινήσεις χρηστών για να ομαλοποιήσουμε το δείγμα θα ήταν λάθος. Έτσι, και χρησιμοποιήσαμε cell arrays, στο Matlab τα οποία και μας έδωσαν την ευελιξία να έχουμε συνεδρίες διαφορετικού πλήθους κινήσεων.

Επίσης θέλαμε η ομαδοποίηση να γίνεται κατευθείαν από το Matlab χωρίς να χρειάζεται να παρέμβουμε εμείς με κάποιο πρόγραμμα επεξεργασίας κειμένου, ή ούτε καν να ανοίξουμε το αρχείο με τα δεδομένα, έτσι εκμεταλλευτήκαμε , τον μοναδικό αριθμό συνεδρίας ώστε να γίνει η ομαδοποίηση. Άρα διαβάζοντας αρχικά το αρχείο το Matlab καταφέραμε να έχουμε τελικά ως αποτέλεσμα την ακολουθία έτοιμη για να τροφοδοτήσουμε το μοντέλο σε μια μεταβλητή με το όνομα seq, όπως και φαίνεται στον κώδικα που χρησιμοποιήσαμε στο Matlab, στο παράρτημα αυτής της εργασίας.

Στην συνέχεια εφόσον είχαμε σε σωστή μορφή την ακολουθία που θα τροφοδοτούσε τον αλγόριθμο, χρειαζόταν να εισαγάγουμε τις αρχικές τιμές των πινάκων A και B, όπως και τους περιγράφουμε στο προηγούμενο κεφάλαιο, καθώς και την επαναληπτική διαδικασία , ώστε

να προκύψει εν τέλει το μοντέλο της κάθε περίπτωσης.

Ο πειραματισμός που έγινε κατά την εκπαίδευση του μοντέλου αποτελείται από διαφορετικές αρχικές τιμές πιθανοτήτων των πινάκων Α και Β, καθώς και από διαφορετικό αριθμό hidden states- και τα δύο απαραίτητα στοιχεία για την μοντελοποίηση με τον αλγόριθμο.

Ο αλγόριθμος που χρησιμοποίησαμε είναι ο Baum-Welch ο οποίος και υλοποιείται από βιβλιοθήκη του Matlab με την εντολή hmmtrain, πληροφορίες για την οποία δίνονται στο παράρτημα.

5.6 Πειραματισμός με διαφορετικές αρχικές τιμές

Όσον αφορά τα hidden states είχαμε από 2, μέχρι και 7, αριθμός ο οποίος καλύπτει όλο το πιθανό φάσμα για βέλτιστο μοντέλο σύμφωνα με την θεωρία των HMM. Το άλλο σημαντικό κομμάτι ήταν οι αρχικές τιμές των πινάκων Α και Β , ο πειραματισμός με τις οποίες οδηγούσε σε διαφορετικά αποτελέσματα. Για να βρούμε το βέλτιστο μοντέλο στην συνέχεια με βάση την αξιολόγηση του με τον αλγοριθμο Forward και την τεχνική του Cross-Validation χρησιμοποιήσαμε τρεις διαφορετικές περιπτώσεις αρχικών τιμών ώστε να προκύψει το βέλτιστο αποτέλεσμα στην συνέχεια. Στην πρώτη περίπτωση επιλέξαμε τιμές για τον πίνακα Α και Β με περίπου των a_{ij} και b_j 1/n και 1/m, στην δεύτερη περίπτωση αρχικές τιμές ακριβώς 1/n και 1/m, και στην τρίτη περίπτωση τυχαίες αρχικές τιμές. Φυσικά σε κάθε περίπτωση πρέπει να προσέξουμε ο αριθμός των επιμέρους αθροισμάτων σε κάθε γραμμή του πίνακα να ισούται με 1 με βάση την θεωρία της στατιστικής και των πινάκων.

Έτσι προκύπτουν οι πίνακες με αρχικές τιμές περίπου 1/n και 1/m, ακριβώς 1/n και 1/m ,καθώς και με τυχαίες αρχικές τιμές ανάλογα με τον αριθμό των hidden states και δίνονται ακριβώς στο παράρτημα.

Έτσι με βάση αυτές τις αρχικές τιμές θα προκύψουν επτά διαφορετικά μοντέλα όσα και τα hidden states ανά περίπτωση, συνολικά δηλαδή 21 διαφορετικά μοντέλα.

5.7 Τρόπος αξιολόγησης των παραγόμενων μοντέλων

Για το κομμάτι της αξιολόγησης τώρα χρησιμοποιήσαμε δύο τρόπους οι οποίοι περιελάμβαναν την χρήση του αλγορίθμου forward με την τεχνική του leave-one-out Cross-Validation με n folds [17].

Αυτό που ουσιαστικά κάναμε με τον πρώτο τρόπο αξιολόγησης, ήταν να υλοποιούμε το μοντέλο με τον Baum-Welch επαναληπτικά τόσες φορές όσες και οι συνεδρίες αλλά με μια συνεδρία λιγότερη κάθε φορά, με την οποία και τρέχαμε στην συνέχεια τον αλγόριθμο Forward για να υπολογίσει την πιθανότητα να προκύψει αυτή. Δηλαδή, κατά την πρώτη επανάληψη, τροφοδοτήσαμε τον αλγόριθμο hmmtrain του Matlab με όλες τις συνεδρίες εκτός από της πρώτη, ώστε να προκύψει ένα μοντέλο (πίνακας A και B) και μετά τρέξαμε τον αλγόριθμο Forward με την συγκεκριμένη ακολουθία και υπολογίσαμε την πιθανότητα που έχει να προκύψει αυτή η ακολουθία. Κατά την δεύτερη επανάληψη ομοίως τροφοδοτήσαμε το μοντέλο με όλες τις ακολουθίες εκτός από την δεύτερη την οποία και κρατήσαμε για να υπολογίσουμε την πιθανότητα της να προκύψει με τον Forward αλγόριθμο. Η διαδικασία αυτή ακολουθήθηκε τόσες φορές όσες και οι ακολουθίες που είχαμε, και για τόσες φορές όσες όλες μας οι δοκιμές, δηλαδή 231 συνεδρίες επί 21 τρόποι άρα 4851 φορές τον αλγόριθμο Forward. Από κάθε έναν από τους 21 τρόπους προέκυπτε και ένας πίνακας 1x231 θέσεων ο οποίος και είχε τις πιθανότητες να προκύψει η κάθε συνεδρία.

Για να εκτελέσουμε τον αλγόριθμο Forward του οποίου ο κώδικας που χρησιμοποιήθηκε δίνεται στο παράρτημα, χρειαζόμασταν το μοντέλο που προέκυπτε κάθε φορά από τον αλγόριθμο Baum-Welch καθώς και τον πίνακα αρχικών καταστάσεων pi ο οποίος περιγράφει την πιθανότητα για το σε ποια κατάσταση θα πάει αρχικά το μοντέλο μας. Για τον τιμή του εκάστοτε πίνακα pi χρησιμοποιήσαμε τιμές 1/n σε κάθε περίπτωση ανάλογα με τον αριθμό των hidden states. Και φυσικά τροφοδοτούσαμε και την ακολουθία που είχαμε αφήσει εκτός της εκπαίδευσης του μοντέλου.

Κατά τον δεύτερο τρόπο αξιολόγησης των παραγόμενων μοντέλων χρησιμοποιήσαμε τον μέσο όρο σε κάθε περίπτωση. Δηλαδή, με ίδιο αριθμό επαναλήψεων 231 φορές επί 21, σε κάθε μία επανάληψη αφήναμε εκτός μια ακολουθία για την εκπαίδευση του μοντέλου, αλλά ο έλεγχος με τον αλγόριθμο Forward γινόταν όχι μόνο με την ακολουθία που αφήναμε εκτός, αλλά με όλες τις ακολουθίες του μοντέλου, με αποτέλεσμα να προκύψουν πίνακες

μεγέθους 231x231 σε κάθε περίπτωση. Στην συνέχεια όμως για να συγκρίνουμε τα αποτελέσματα με τον πρώτο τρόπο ελέγχου, παίρνουμε την μέση τιμή της κάθε στήλης.

Κεφάλαιο 6

Αξιολόγηση των αποτελεσμάτων

6.1 Αποτελέσματα των πειραμάτων του μοντέλου μας

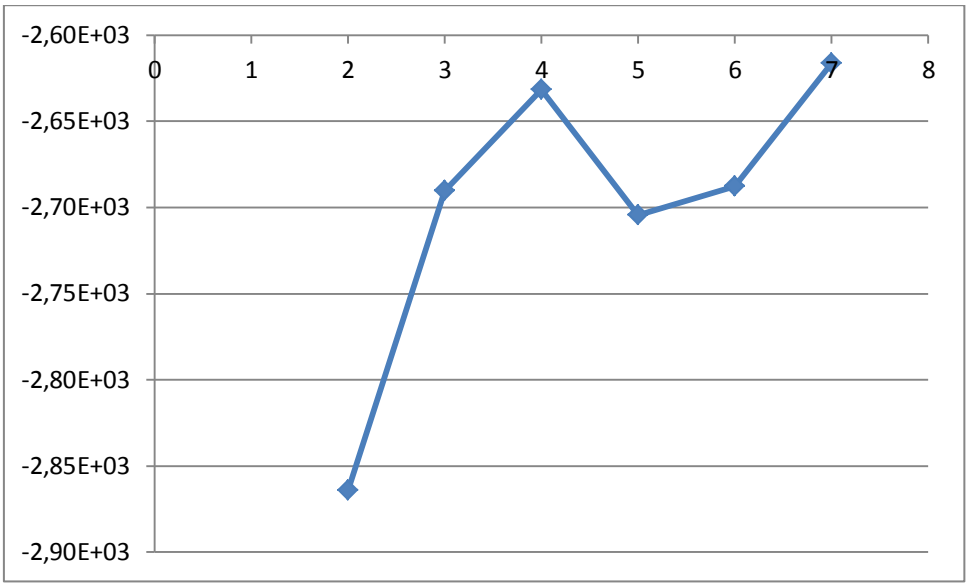
Σε αυτό το σημείο έχοντας δώσει όλο το απαραίτητο θεωρητικό και πρακτικό υπόβαθρο για όλη την διαδικασία την οποία ακολουθήσαμε, θα παραθέσουμε τα αποτελέσματα της μελέτης, η οποία διεξήχθη με 10 χρήστες οι οποίοι χρησιμοποιούσαν για διάστημα ενός μήνα την εφαρμογή CRUISE και κάνανε αναζητήσεις σε αυτήν. Έχοντας πλέον μαζέψει αρκετό υλικό από 231 συνδερίες τους, το οποίο αποτελείται από περισσότερες των 2000 κινήσεων, κρατήσαμε τα δεδομένα στην βάση δεδομένων MySql με την μέθοδο που περιγράψαμε στα προηγούμενα κεφάλαια, και ακολουθώντας τα βήματα που δίνονται στο κομμάτι της υλοποίησης προκύπτουν τα αποτελέσματα που θα δωθούν παρακάτω.

Εφόσον πλεόν έχουμε όλα τα αποτελέσματα καταχωρημένα σε πίνακες (μεταβλητές) , πρέπει να βγάλουμε και τα συμπεράσματα για το πιο μοντέλο έιναι βέλτιστο στην δικιά μας εφαρμογή. Για να προκύψουν αυτά τα συμπεράσματα πρέπει να συγκρίνουμε τα αποτελέσματα και ο καταλληλότερος τρόπος είναι με την βοήθεια της λογαριθμικής πιθανότητας των αθροισμάτων τους (Log Likelihood) τα οποία και υπολογίζουμε με την βοήθεια του Matlab.

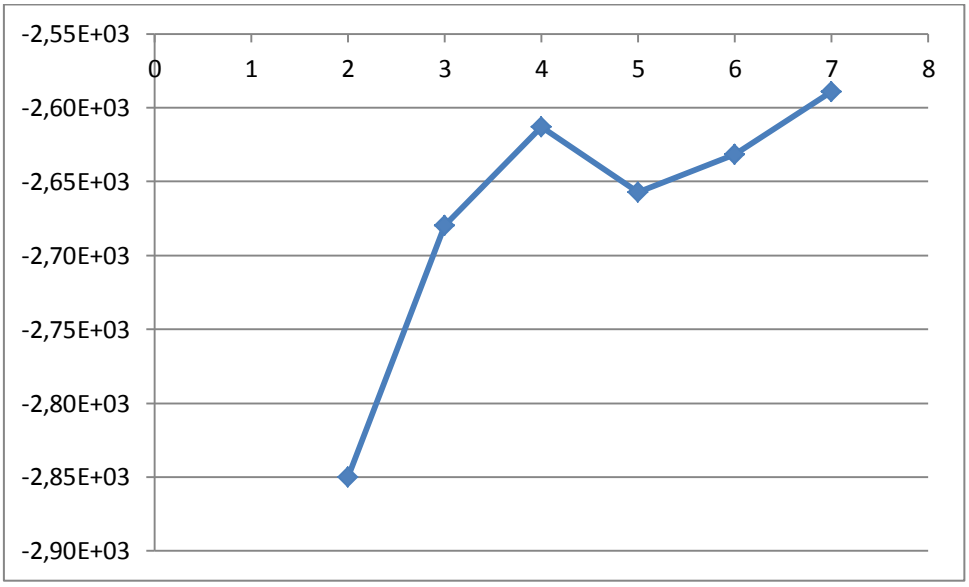
Έχοντας πλέον κάνει τις δοκιμές αυτές , και έχοντας επεξεργαστεί τα δεδομένα όπως περιγράψαμε παραπάνω, παραθέτουμε παρακάτω τα αποτελέσματα τους.

	approximately 1/n 1/m		exactly 1/n 1/m		random starting values	
Hidden States	one time forward	n times forward	one time forward	n times forward	one time forward	n times forward
2	-2,86E+03	-2,85E+03	-2,84E+03	-2,83E+03	-2,72E+03	-2,71E+03
3	-2,69E+03	-2,68E+03	-2,73E+03	-2,71E+03	-2,76E+03	-2,72E+03
4	-2,63E+03	-2,61E+03	-2,84E+03	-2,83E+03	-2,67E+03	-2,64E+03
5	-2,70E+03	-2,66E+03	-2,84E+03	-2,83E+03	-2,70E+03	-2,65E+03
6	-2,69E+03	-2,63E+03	-2,84E+03	-2,83E+03	-2,72E+03	-2,66E+03
7	-2,62E+03	-2,59E+03	-2,66E+03	-2,63E+03	-2,64E+03	-2,60E+03

Επίσης παραθέτουμε στην συνέχεια και τα διαγράμματα [17] τα οποία δείχνουν τα παραπάνω αποτελέσματα ώστε να μας διευκολύνουν να βγάλουμε συμπεράσματα.



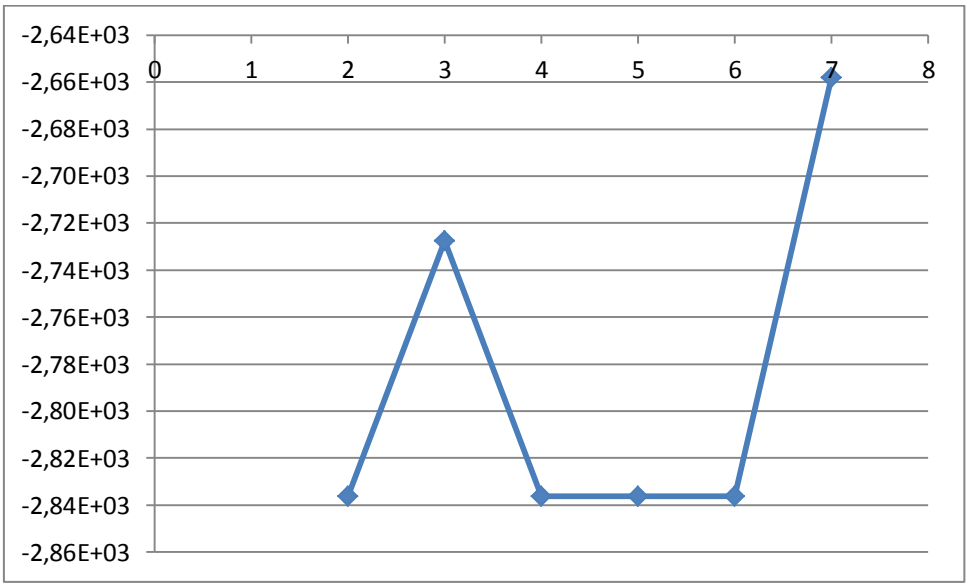
Αρχικές τιμές με περίπου 1/ν,1/μ , μια φορά εκτέλεση του αλγορίθμου Forward



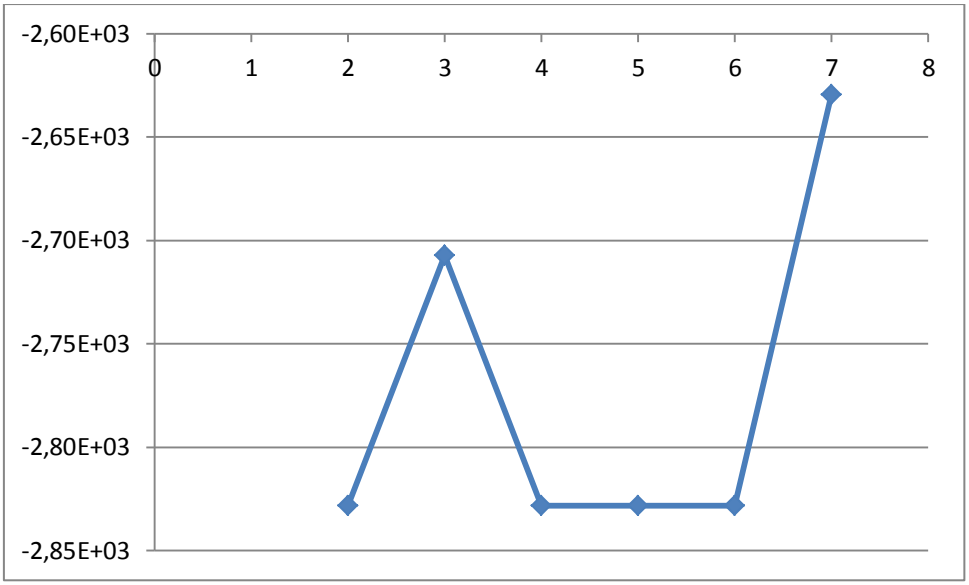
Αρχικές τιμές με περίπου 1/ν,1/μ , 231 φορές εκτέλεση του αλγορίθμου Forward

Στα αποτελέσματα μας του μοντέλου με αρχικές τιμές περίπου 1/ν και 1/μ, και στις δύο περιπτώσεις , δηλαδή και στην περίπτωση που αξιολογήσαμε τα αποτελέσματα υλοποιώντας μια φορά τον αλγόριθμο Forward κατά την διαδικασία του Cross-Validation καθώς και στην περίπτωση που υλοποίησαμε 231 φορές τον αλγόριθμο Forward παρατηρούμε πως τα αποτελέσματα παρουσιάζουν τα ίδια μέγιστα δηλαδή στα μοντέλα με τέσσερα και με επτά καταστάσεις. Φυσικά, το μοντέλο το οποίο έχει την μεγαλύτερη τιμή είναι το βέλτιστο για την περίπτωση μας, καθώς παρουσιάζει την μεγαλύτερη πιθανότητα

να συμβεί. Μια άλλη παρατήρηση ανάλογα τις καταστάσεις που έχουμε είναι πως το μοντέλο με δύο καταστάσεις είναι η χειρότερη αναπαράσταση του μοντέλου μας, και φαίνεται πως οι δύο καταστάσεις δεν αρκούν ώστε να περιγράψουν όλες τις πνευματικές καταστάσεις που βρίσκονται οι χρήστες. Στην συνέχεια βλέπουμε πως όσο αυξάνονται οι καταστάσεις βελτιώνεται και το μοντέλο που μας παρουσιάζουν μέχρι τις τέσσερις καταστάσεις. Μετά τις τέσσερις καταστάσεις παρατηρούμε πως υπάρχει μια πτώση στα μοντέλα που δείχνουν και τελικά υπάρχει ανάκαμψη πάλι στις επτά καταστάσεις όπου και έχουμε τα βέλτιστα αποτελέσματα. Άρα είναι εμφανές πως γενικά το μοντέλο μας δεν μπορεί να περιγραφεί με λιγότερες από τέσσερις καταστάσεις ικανοποιητικά. Επίσης, θεωρούμε πως δεν υπάρχει λόγος να ασχοληθούμε με περισσότερες των επτά καταστάσεων, αφού μετά ενώ ίσως θα έχουμε ελαφρά καλύτερα αποτελέσματα, οι καταστάσεις θα αρχίσουν να επαναλαμβάνονται απλώς , δείχνοτας τα ίδια αποτελέσματα με μικρές διαφοροποιήσεις.

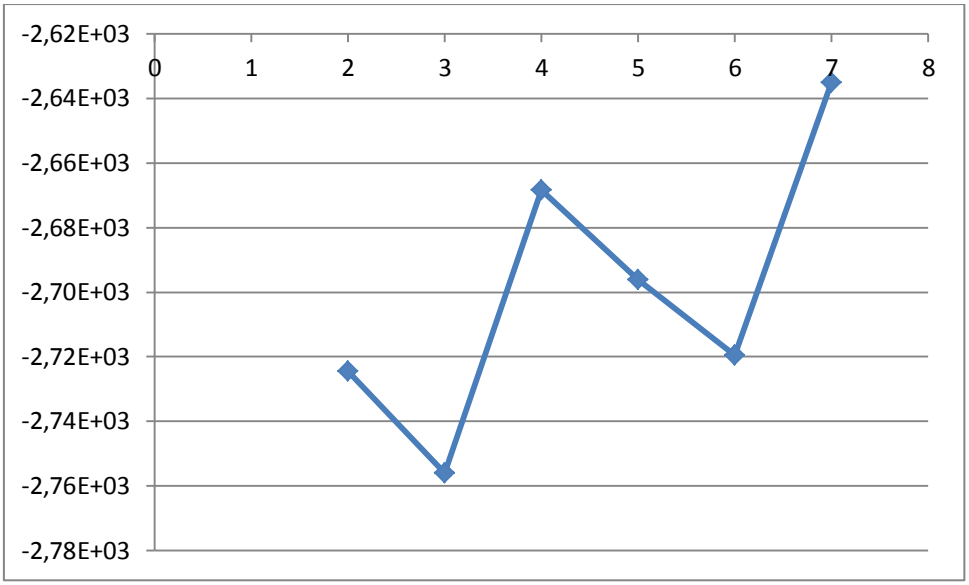


Αρχικές τιμές με ακριβώς 1/ν,1/μ , μια φορά εκτέλεση του αλγορίθμου Forward

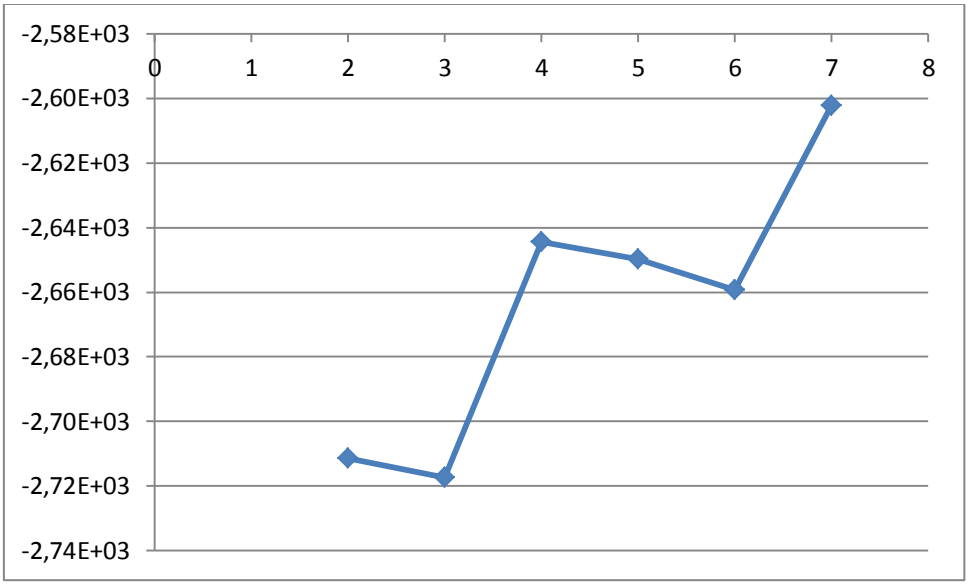


Αρχικές τιμές με ακριβώς 1/ν,1/μ , 231 φορές εκτέλεση του αλγορίθμου Forward

Στα αποτελέσματα μας του μοντέλου με αρχικές τιμές ακριβώς 1/ν και 1/μ, και στις δύο περιπτώσεις αξιολόγησης είτε υλοποιώντας μια φορά τον αλγόριθμο Forward κατά την διαδικασία του Cross-Validation είτε υλοποιώντας τον 231 φορές παρατηρούμε πως τα αποτελέσματα παρουσιάζουν τα ίδια μέγιστα δηλαδή στα μοντέλα με τρεις και με επτά καταστάσεις. Επίσης το ότι παίρνουμε πάλι όπως στην προηγούμενη περίπτωση σχεδόν ίδια αποτελέσματα με αμελητέες διαφορές βλέπουμε πως και οι δύο τρόποι του Cross-Validation συγκλίνουν στα ίδια, άρα και η αξιολόγηση μέσω αυτής της μεθόδου κρίνεται ικανοποιητική. Μια άλλη παρατήρηση ανάλογα τις καταστάσεις που έχουμε είναι πως το μοντέλο με δύο καταστάσεις καθώς και τα μοντέλα με τις τέσσερις, πέντε καθώς και έξι καταστάσεις παρουσιάζουν τις χειρότερες περιπτώσεις. Επίσης, θεωρούμε πως δεν υπάρχει λόγος να ασχοληθούμε με περισσότερες των επτά καταστάσεων, για τον ίδιο λόγο που αναφέραμε και προηγούμενως καθώς δεν μπορούν να προκύψουν περισσότερες των επτά καταστάσεων χωρίς να επαναλαμβάνονται στο πρόβλημα μας. Μέχρι εδώ, δηλαδή αναλύοντας τα δύο μοντέλα με διαφορετικές αρχικές τιμές, παρατηρούμε πως έχουμε απόκλιση όσον αφορά τα μοντέλα με τις τρεις και τις τέσσερις καταστάσεις καθώς διαφορετικά αποτελέσματα παίρνουμε, όμως έχουμε απόλυτη συμφωνία πως το μοντέλο με τις επτά καταστάσεις είναι το βέλτιστο.



Τυχαίες αρχικές τιμές , μια φορά εκτέλεση του αλγορίθμου Forward



Τυχαίες αρχικές τιμές , 231 φορές εκτέλεση του αλγορίθμου Forward

Στα αποτελέσματα του μοντέλου με τυχαίες αρχικές τιμές, και στις δύο περιπτώσεις αξιολόγησης είτε υλοποιώντας μια φορά τον αλγόριθμο Forward κατά την διαδικασία του Cross-Validation είτε υλοποιώντας τον 231 φορές παρατηρούμε πως τα αποτελέσματα παρουσιάζουν την ίδια πορεία με τα ίδια μέγιστα δηλαδή στα μοντέλα με τέσσερις και με επτά καταστάσεις και τις ίδιες πορείες στα υπόλοιπα. Παρατηρούμε επίσης πως τα

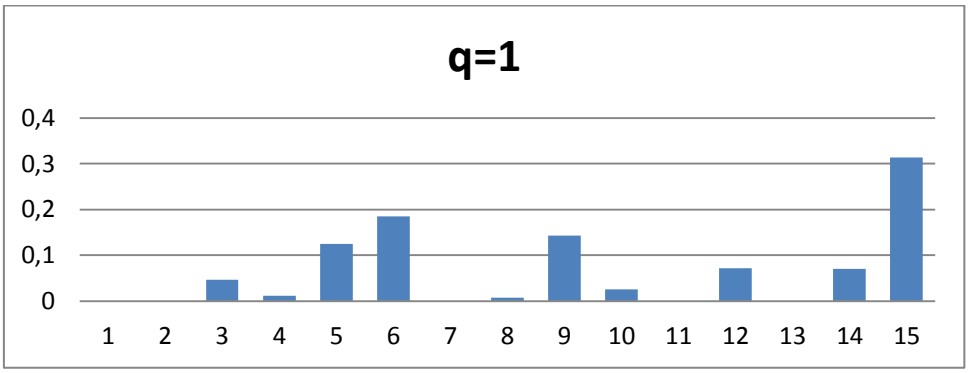
αποτελέσματα που προκύπτουν από αυτό το μοντέλο συμβαδίζουν αρκετά με τα αποτελέσματα που πήραμε από την ανάλυση των άλλων δύο, και συγκεκριμένα μοιάζουν περισσότερο με το μοντέλο με αρχικές τιμές περίπου 1/ν και 1/μ, που αναλύσαμε πρώτο.

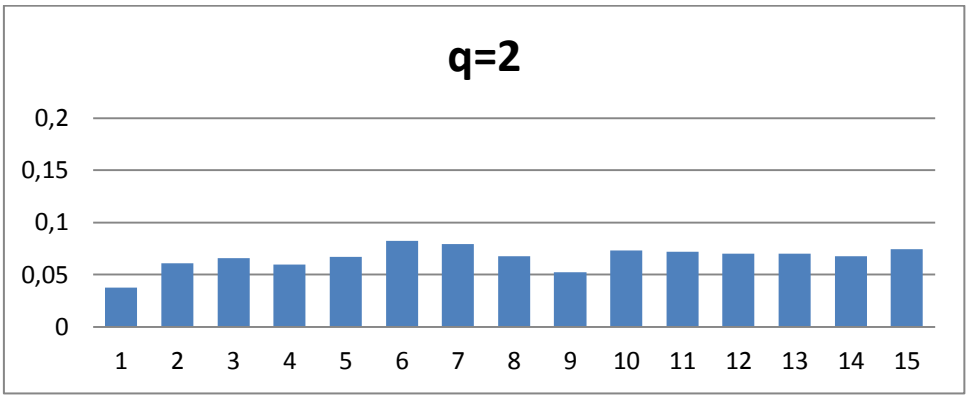
Από όλα τα παραπάνω που είδαμε και αναλύσαμε παρατηρούμε ότι τα καλυτερα αποτελέσματα προκύπτουν με την χρήση τεσσάρων και επτά κρυμμένων καταστάσεων (Hidden States) όμως με την χρήση των επτά κρυμμένων καταστάσεων παρουσιάζεται το πρόβλημα της ύπαρξης ίδιων καταστάσεων των χρηστών(overfitting) και με περαιτέρω ανάλυση που θα δούμε παρακάτω αυτό θα γίνει εμφανές. Επίσης παρατηρούμε πως με αρχικές τιμές περίπου 1/ν και 1/μ επιτυγχάνουμε πολύ καλές προσεγγίσεις με τις υψηλότερες τιμές πιθανότητας στο βέλτιστο μοντέλο με τις τέσσερις καταστάσεις. Βλέπουμε πως ο αριθμός των hidden states είναι πολύ σημαντικός παράγοντας και καθοριστικός για την ακρίβεια του βέλτιστου μοντέλου όπως γίνεται φανερό από τα παραπάνω, καθώς η αλλαγή των hidden states συμβάλει δραματικά στην αλλαγή των αποτελεσμάτων. Επίσης, εκτός από τα Hidden States, είναι φανερό πως και οι αρχικές τιμές του μοντέλου παίζουν πολύ σημαντικό παράγοντα στα τελικά αποτελέσματα, ειδικά σε μοντέλα σαν και αυτό που μελετάμε και δεν υπάρχουν πληροφορίες εκ των προτέρων. Έτσι ήμασταν αναγκασμένοι να δοκιμάσουμε διάφορες αρχικές τιμές, ώστε να καταλήξουμε στις καλύτερες για το μοντέλο μας.

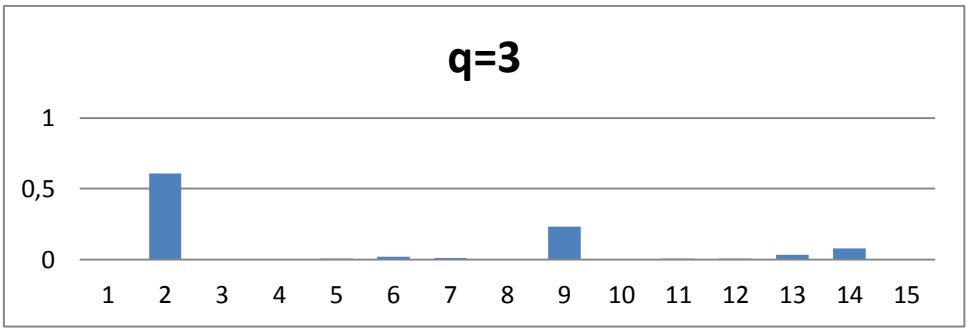
6.2 Ανάλυση και διαγράμματα των αποτελεσμάτων

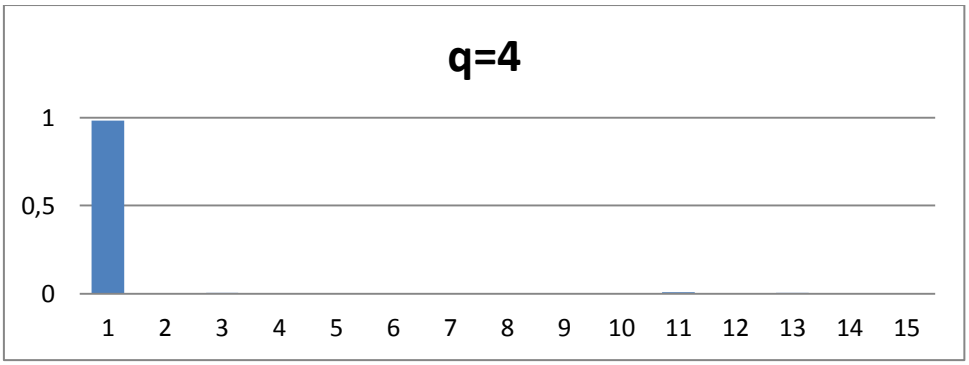
Επίσης, από τα μοντέλα που προέκυψαν για τα hidden states με τα βέλτιστα αποτελέσματα δηλαδή με τέσσερα και επτά hidden states και αρχικά θα δούμε το πρόβλημα της υπερκάλυψης (overfit) αναλύοντας παρακάτω τα αποτελέσματα του μοντέλου των επτά καταστάσεων, έχουμε τις παρακάτω πνευματικές καταστάσεις των χρηστών [17] (emission plots) με αρχικές τιμές περίπου 1/n,1/m , σύμφωνα με τις τιμές του παρακάτω πίνακα:

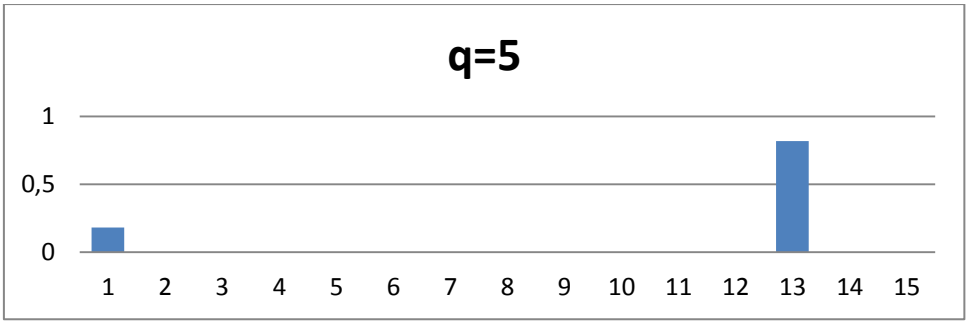
q/ id	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1,05E -13	4,30E -209	0,046 585	0,011 51	0,125 054	0,185 204	8,17E -169	0,007 499	0,142 611	0,026 037	2,78E -114	0,071 855	3,58E -25	0,070 178	0,313 466
2	0,037 277	0,061 06	0,065 924	0,059 823	0,066 789	0,082 201	0,079 059	0,067 731	0,052 42	0,073 065	0,071 686	0,070 405	0,070 176	0,067 951	0,074 432
3	6,59E -122	0,609 985	3,07E -108	1,14E -156	0,005 26	0,021 594	0,013 235	8,88E -259	0,231 122	4,10E -68	0,005 178	0,001 783	0,032 457	0,079 385	6,63E -88
4	0,981 314	3,39E -140	0,004 301	2,97E -243	6,44E -258	6,47E -67	0	1,70E -271	1,59E -21	6,60E -202	0,009 293	3,35E -188	0,005 092	0	5,18E -131
5	0,184 118	9,75E -38	8,07E -49	2,57E -230	3,30E -175	6,04E -29	4,04E -248	4,86E -272	2,80E -72	1,46E -101	4,09E -135	1,15E -52	0,815 882	2,89E -129	1,14E -41
6	0,981 314	3,39E -140	0,004 301	2,97E -243	6,44E -258	6,47E -67	0	1,70E -271	1,59E -21	6,60E -202	0,009 293	3,35E -188	0,005 092	0	5,18E -131
7	0,981 314	3,39E -140	0,004 301	2,97E -243	6,44E -258	6,47E -67	0	1,70E -271	1,59E -21	6,60E -202	0,009 293	3,35E -188	0,005 092	0	5,18E -131

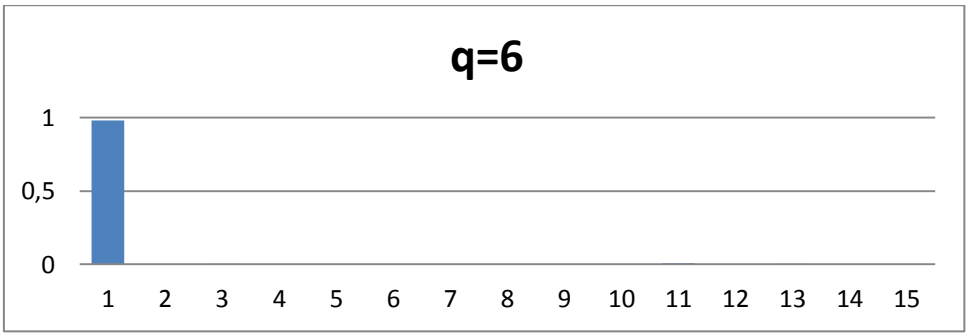

q=1
0,4
0,3
0,2
0,1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

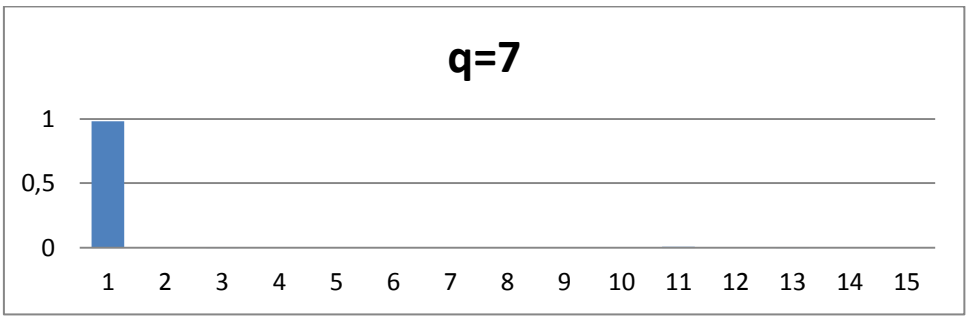

q=2
0,2
0,15
0,1
0,05
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15


q=3
1
0,5
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15


q=4
1
0,5
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15







Κάποιες πολύ σημαντικές παρατηρήσεις που προκύπτουν από τα παραπάνω είναι οι συσχετίσεις ανάμεσα σε κάποιες κινήσεις των χρηστών, οι οποίες φανερώνουν την κατάσταση που βρίσκεται η αναζήτηση τους. Συγκεκριμένα στο q=1 παρατηρούμε ότι η επιλογή Ανακάτεμα(Shuffle) συνδέεται πολύ στενά με τις κινήσεις των χρηστών 5,6 και 9 καθώς και σε μικρότερο βαθμό και με άλλες, οι οποίες όμως κινήσεις είναι σχετικές με έρευνα κατά την αναζήτηση του, δηλαδή κινήσεις για να δει πληροφορίες, ή να προσθέσει κάποιον όρο στην λίστα του, ή να δει μια εικόνα. Αυτό μας δείχνει πληροφορίες ότι μετά από αυτές τις κινήσεις ο χρήστης πιθανόν και να μην βρίσκει επιπλέον χρήσιμα πράγματα με τον τρόπο που του εμφανίζονται, και έτσι επιλέγει να κάνει Ανακάτεμα της μορφής που εμφανίζονται ώστε να βρει κάτι ακόμα σχετικά το οποίο θα το χρησιμοποιήσει με τις ίδιες κινήσεις. Κατά το q=2 δεν μπορούμε να βγάλουμε κάποια συμπεράσματα αφού τα αποτελέσματα είναι αρκετά ομαλά και δεν μπορούμε να κάνουμε κάποια συσχέτιση των κινήσεων.

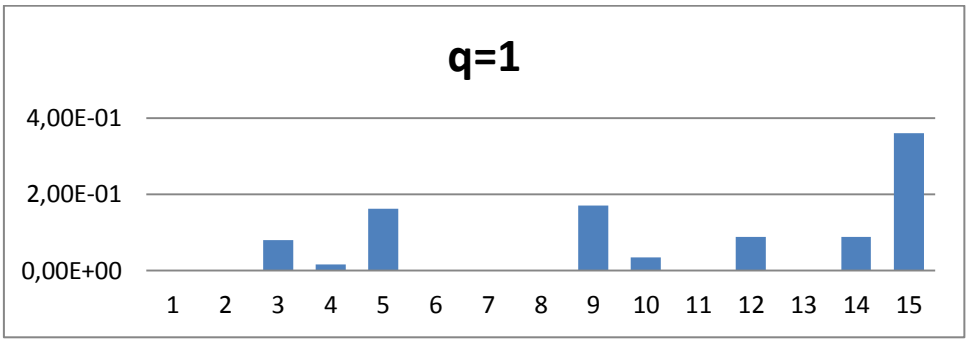
Στην συνέχεια κατά την εικόνα q=3 βλέπουμε πως υπάρχει άμεση σχέση των κινήσεων 2 και 9 καθώς και σε μικρότερο βαθμό της κίνησης 14. Η κίνηση 2 του χρήστη είναι κλικ με τον κέρσορα σε κάποιον όρο κειμένου του Twitter για περισσότερες πληροφορίες και η κίνηση 9 είναι προσθήκη του ενός όρου κειμένου του Twitter στην λίστα του, ενώ η κίνηση 14 είναι η νέα αναζήτηση χρησιμοποιώντας τα δεδομένα που έχει ήδη προσθέσει στην λίστα του. Όπως γίνεται εμφανές και αυτή η πληροφορία είναι πολύ χρήσιμη, καθώς βλέπουμε άμεσα πως συνήθεια του χρήστη είναι αφού κάνει κλικ σε έναν όρο για περισσότερες πληροφορίες, αμέσως μετά να τον προσθέτει στην λίστα του για την συνέχεια της αναζήτησης του. Δηλαδή φαίνεται πως η πληροφορία αυτή του ήταν χρήσιμη στην αναζήτηση και την αξιοποίησε αμέσως κάνοντας στην συνέχεια αναζήτηση με βάση τα δεδομένα αυτά. Και το ότι υπάρχει μικρότερος βαθμός αναζητήσεων μας δείχνει ότι ο χρήστης μαζεύει στην λίστα του περισσότερες της μια πληροφορίας πριν προχωρήσει φυσικά σε αναζήτηση εκ νέου. Αν αντιθέτως υπήρχε συσχέτιση της κίνησης 2 με κάποια άλλη κίνηση όπως το να είναι ανενεργός (idle) ή να κάνει νέα αναζήτηση, θα ήταν δείγμα ότι έχει κορεστεί η υπάρχουσα αναζήτηση του και έχει χάσει ήδη το ενδιαφέρον του.

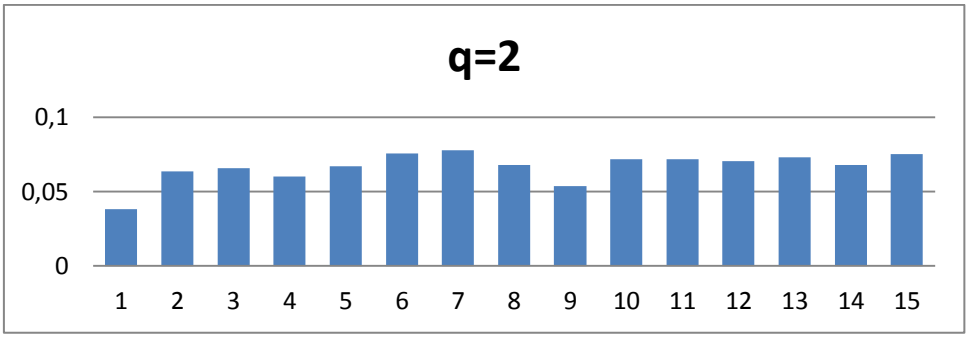
Κατά την κατάσταση της εικόνας q=4 δεν μπορούμε να κάνουμε κάποια συσχέτιση μεταξύ των κινήσεων, βλέπουμε όμως ότι ο χρήστης είναι στην κατάσταση συνεχώς νέων αναζητήσεων. Δηλαδή είναι στην αρχή της διερευνητικής αναζήτησης και ψάχνει γενικά να βρει πληροφορίες οι οποίες θα τον ενδιαφέρουν και θα εκμεταλλευτεί στην συνέχεια. Αυτό βέβαια μπορεί να έχει είτε θετικό είτε αρνητικό αντίκτυπο στο πόσο αποδοτική είναι η εφαρμογή μας, καθώς οι συνεχώς νέες αναζητήσεις έχουν διφορούμενα αποτελέσματα, όμως αυτό δεν θα το αναλύσουμε στο κομμάτι αυτό που μας ενδιαφέρουν μονάχα οι πνευματικές καταστάσεις των χρηστών και πως μπορούμε να τις μελετήσουμε και να τις αξιοποιήσουμε. Η ίδια αξιολόγηση προκύπτει και από τις καταστάσεις των χρηστών κατά τις εικόνες q=6 και q=7 όπου ο χρήστης βρίσκεται στην ίδια κατάσταση αναζήτησης.

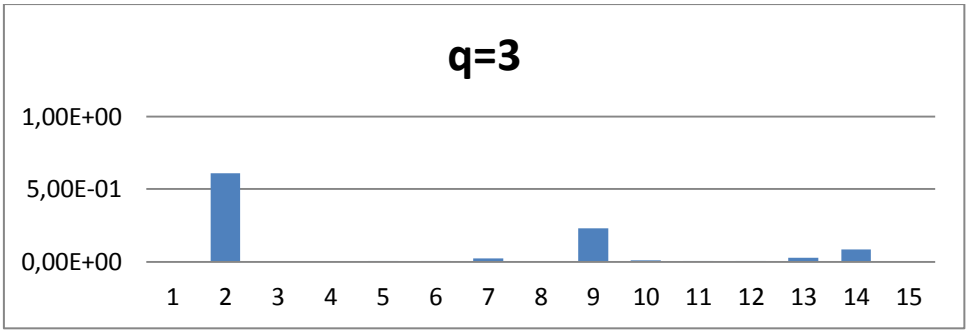
Τέλος κατά την εικόνα q=5 βλέπουμε πως υπάρχει συσχέτιση μεταξύ των κινήσεων 1 και 13 δηλαδή της νέας αναζήτησης με την αδράνεια(idle) κάτι που μας δείχνει κατάσταση που ο χρήστης έχει χάσει το ενδιαφέρον του για επιπλέον αναζήτηση, και ίσως έχει ήδη φύγει από την εφαρμογή.

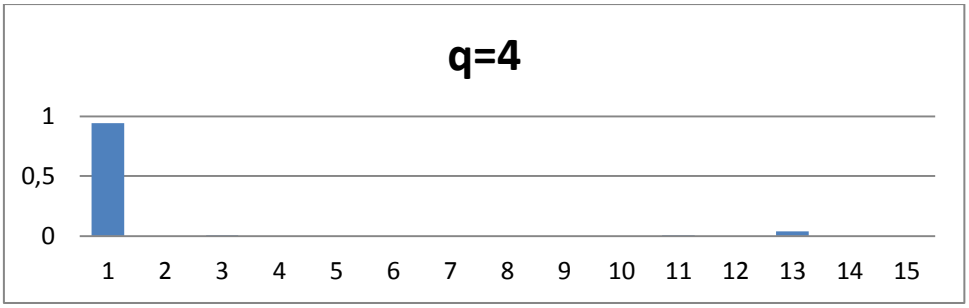
Τα παρακάτω emission plots με αρχικές τιμές ακριβώς 1/n,1/m :

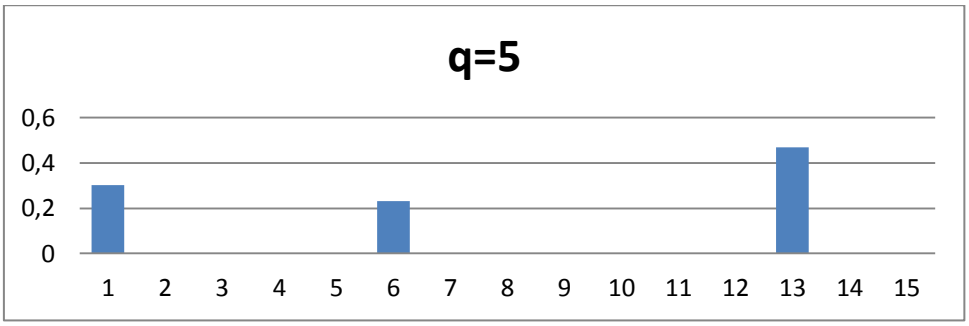
q/id	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	2,08E-33	1,40E-136	0,079491	0,015487	0,161755	5,18E-32	2,33E-146	1,18E-05	0,170398	0,035009	5,17E-93	0,088343	0,000957	0,088189	0,360359
2	0,037995	0,063622	0,065876	0,059969	0,067142	0,075568	0,077782	0,06802	0,053677	0,071757	0,071855	0,070435	0,073241	0,068007	0,075054
3	5,01E-116	0,610259	4,00E-108	1,94E-173	0,004075	0,00545	0,022515	3,38E-272	0,23068	0,008912	0,006217	0,002946	0,026848	0,082098	1,43E-122
4	0,945381	2,03E-139	0,004298	9,69E-248	2,32E-276	7,25E-167	0	1,60E-269	1,59E-12	5,05E-233	0,008677	1,15E-252	0,041645	0	1,48E-137
5	0,301775	4,79E-103	4,73E-111	2,12E-210	2,79E-174	0,230693	1,22E-222	7,64E-298	2,52E-59	2,67E-33	2,08E-110	9,56E-69	0,467532	1,13E-198	2,21E-83
6	0,301775	4,79E-103	4,73E-111	2,12E-210	2,79E-174	0,230693	1,22E-222	7,64E-298	2,52E-59	2,67E-33	2,08E-110	9,56E-69	0,467532	1,13E-198	2,21E-83
7	0,301775	4,79E-103	4,73E-111	2,12E-210	2,79E-174	0,230693	1,22E-222	7,64E-298	2,52E-59	2,67E-33	2,08E-110	9,56E-69	0,467532	1,13E-198	2,21E-83

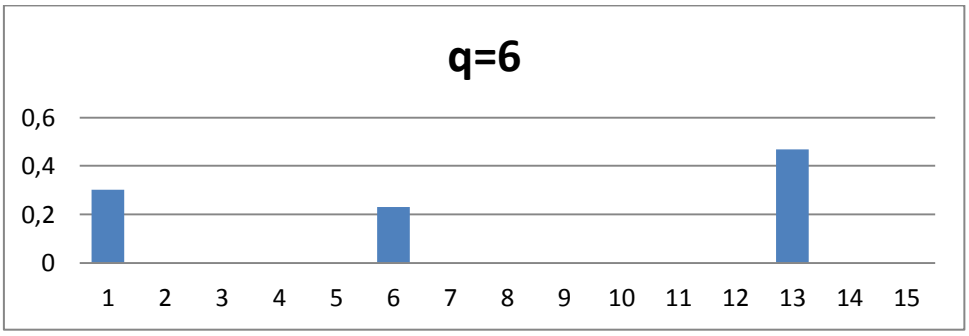


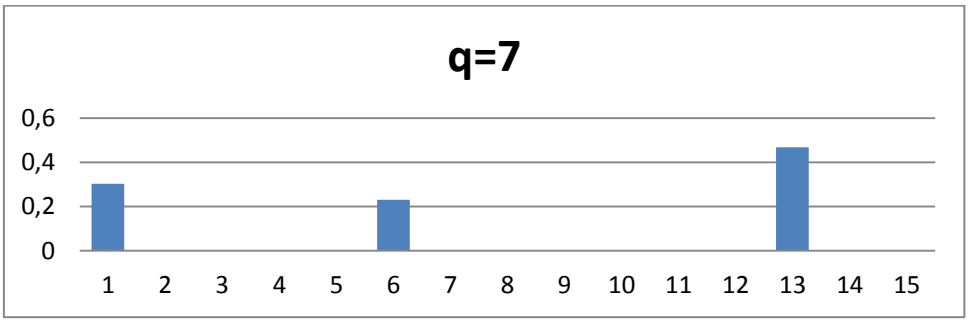












Από τα παραπάνω τώρα προκύπτουν παρόμοια συμπεράσματα όσον αφορά την κατάσταση που βρίσκονται οι χρήστες. Για την πνευματική κατάσταση των χρηστών στην εικόνα με q=1 , παρομοίως με τα προηγούμενα διαγράμματα όπου είχαμε περίπου 1/ν και 1/μ αρχικές τιμές, βλέπουμε πως υπάρχει συσχέτιση στους χρήστες μεταξύ των κινήσεων 5,9 και 15, καθώς και κάποιων άλλων σε μικρότερο όμως βαθμό όπως 3,12,14. Οι βασικές κινήσεις 5,9 και 15 είναι αντίστοιχα κλικ σε σύνδεσμο για περισσότερες πληροφορίες, προσθήκη κάποιου όρου στην λίστα του χρήστη και ανακάτεμα αντίστοιχα. Και οι υπόλοιπες 3,12,14

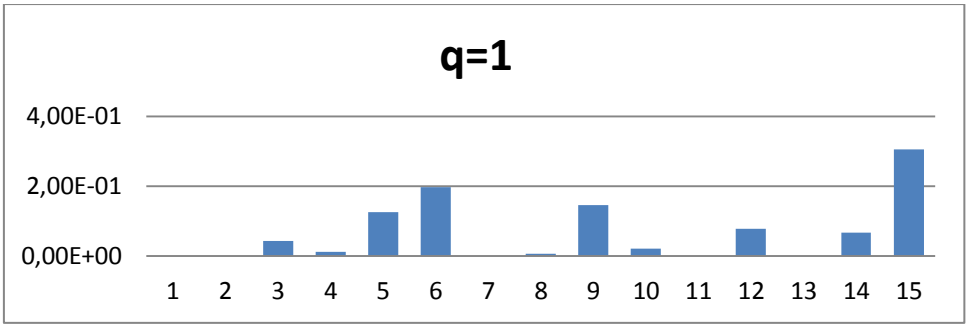
είναι κλικ σε εικόνα για περισσότερες πληροφορίες, προσθήκη συνδέσμου στη λίστα του, καθώς και νέα αναζήτηση με την χρήση των όρων στην λίστα αντιστοίχως. Βλέπουμε δηλαδή με βάση την συσχέτιση των κινήσεων αυτών πως ο χρήστης κάνει συγκεκριμένη αναζήτηση σε συνδέσμους κυρίως και λιγότερο σε εικόνες και όρους του Twitter. Δηλαδή μπορούμε με βάση τις πληροφορίες αυτές να έχουμε ακριβή εικόνα για τις σχετικότερες κινήσεις του χρήστη κατά την κατάσταση αυτή.

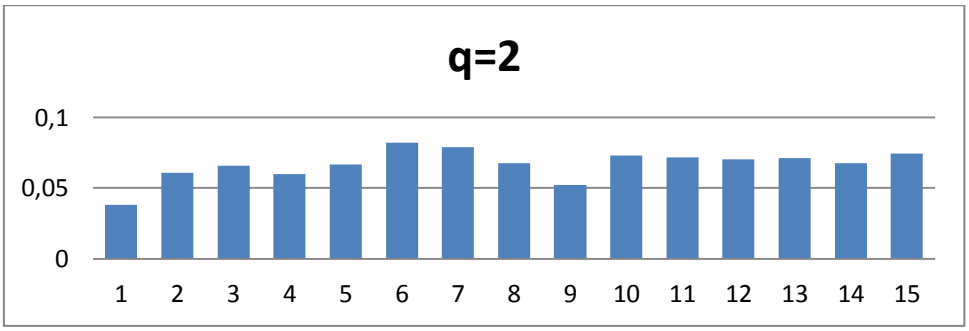
Στην συνέχεια η αξιολόγηση που προκύπτει από τις εικόνες q=2, q=3 και q=4 είναι η ίδια που κάναναμε και με τα προηγούμενα διαγράμματα, δηλαδή νέες αναζητήσεις καθώς και συσχέτιση ανάμεσα σε κλικ για πληροφορίες και προσθήκη στην λίστα.

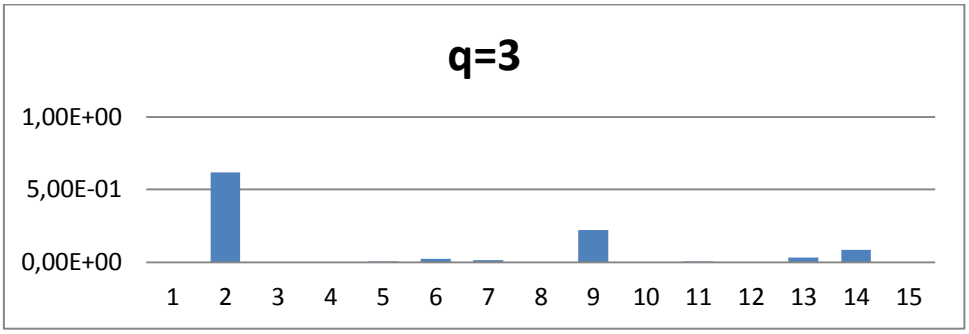
Τέλος κατά τις πνευματικές καταστάσεις των εικόνων q=5, q=6 και q=7 βλέπουμε να υπάρχει μεγάλη συσχέτιση των κινήσεων του χρήστη, 1,6 και 13. Αυτές οι κινήσεις είναι η νέα αρχική αναζήτηση, η επιλογή μιας εικόνας για προβολή με τον κέρσορα του ποντικιού, καθώς και η αδράνεια αντίστοιχα. Αυτό μπορεί να συσχετιστεί ότι ο χρήστης είναι σε μια κατάσταση που δεν ενδιαφέρεται να κάνει διερευνητική αναζήτηση αλλά κάνει απλές αναζητήσεις εικόνων οι οποίες ίσως δεν τον ενδιαφέρουν τόσο ως προς το κομμάτι της αναζήτησης του. Είναι και ένας τρόπος επίσης να τελειώσει την συνεδρία του, κάνοντας κάποιες τελευταίες αναζητήσεις εικόνων πριν αφήσει την εφαρμογή.

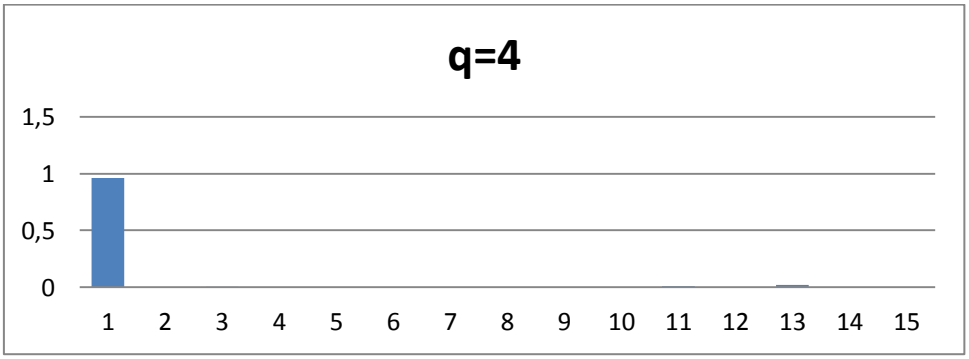
Τα παρακάτω emission plots με τυχαίες αρχικές τιμές:

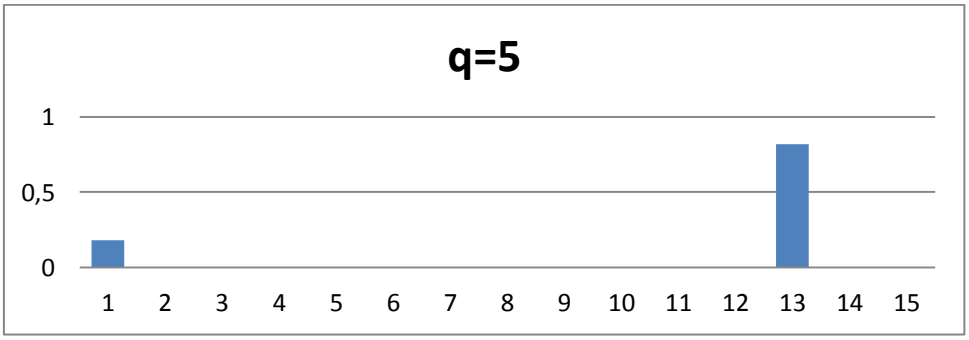
q/id	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	4,06E-22	4,62E-111	0,043332	0,011586	0,125354	0,196684	3,34E-89	0,006054	0,146087	0,02039	3,44E-58	0,078383	3,20E-10	0,06617	0,30596
2	0,038036	0,060729	0,065835	0,059705	0,066646	0,082074	0,078815	0,067619	0,052313	0,073007	0,071473	0,070487	0,071081	0,067724	0,074459
3	7,39E-81	0,618044	4,35E-74	1,01E-96	0,005683	0,021493	0,014328	4,68E-174	0,219305	3,95E-37	0,005394	1,78E-09	0,032001	0,083751	8,46E-67
4	0,961694	1,30E-122	0,005173	5,50E-159	1,58E-159	6,04E-29	1,02E-212	2,33E-189	1,56E-75	1,88E-159	0,011462	5,26E-154	0,021671	2,53E-211	2,24E-109
5	0,181289	2,60E-40	2,22E-35	1,32E-165	3,00E-136	6,36E-08	2,93E-179	2,61E-194	8,21E-66	5,40E-88	6,13E-90	7,33E-62	0,818711	1,58E-101	3,03E-49
6	0,961694	1,30E-122	0,005173	5,50E-159	1,58E-159	6,04E-29	1,02E-212	2,33E-189	1,56E-75	1,88E-159	0,011462	5,26E-154	0,021671	2,53E-211	2,24E-109
7	0,794306	0,09646	1,82E-206	4,34E-197	1,77E-275	1,59E-62	4,72E-301	5,45E-155	0,109234	2,18E-101	2,03E-24	9,89E-69	6,65E-88	3,00E-289	9,53E-78

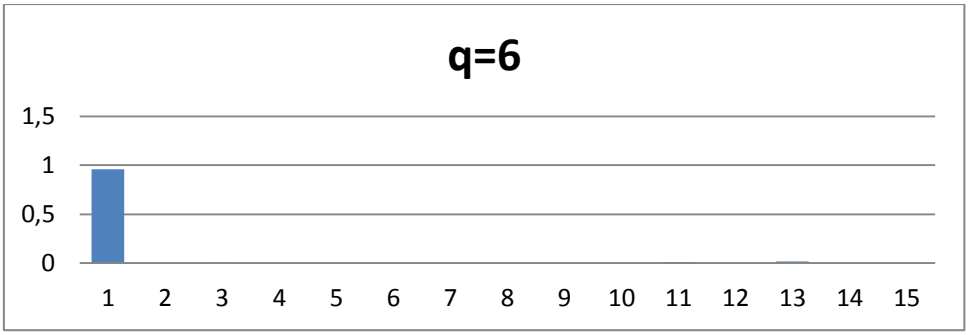

q=1
4,00E-01
2,00E-01
0,00E+00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

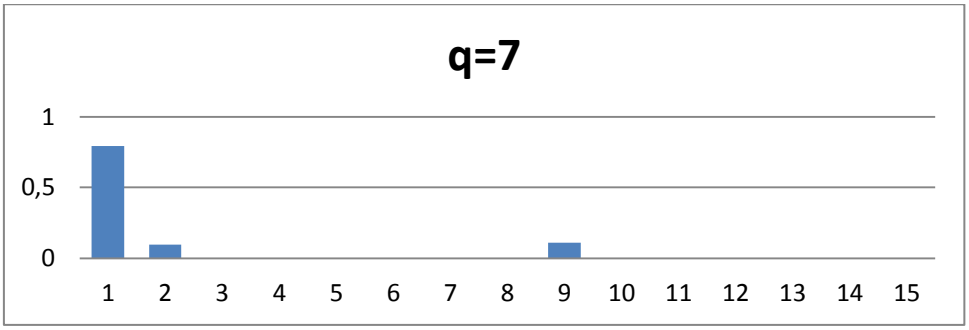

q=2
0,1
0,05
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15


q=3
1,00E+00
5,00E-01
0,00E+00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15


q=4
1,5
1
0,5
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15


q=5
1
0,5
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15





Κατά τα διαγράμματα αυτά βλέπουμε παρόμοιες πληροφορίες με αυτές που έχουμε ήδη αναλύσει στις δύο προηγούμενες περιπτώσεις όπου είχαμε αρχικές τιμές περίπου 1/ν , 1/μ και ακριβώς 1/ν , 1/μ.

Βλέπουμε δηλαδή πως αρχικά στην κατάσταση q=1 έχουμε την ίδια κατάσταση που περιγράψαμε προηγουμένως και προκύπτουν τα ίδια συμπεράσματα, με αμελητέες διαφοροποιήσεις στα μεγάθη των κινήσεων του χρήστη, χωρίς όμως να αλλάζουν την γενική εικόνα για την κατάσταση που βρίσκεται.

Στην κατάσταση q=2 ομοίως οι κινήσεις του χρήστη χωρίζονται ομαλά, με μικρές εναλλαγές μονάχα.

Στην κατάσταση q=3 έχουμε την κατάσταση της συγκεκριμένης αναζήτησης από τον χρήστη σε όρους του Twitter που τους χρησιμοποιεί , ερευνόντας τις πληροφορίες που έχουν, καθώς και προσθέτει αυτούς που τον ενδιαφέρουν περισσότερο στην λίστα του, και συνεχίζει με νέα αναζήτηση.

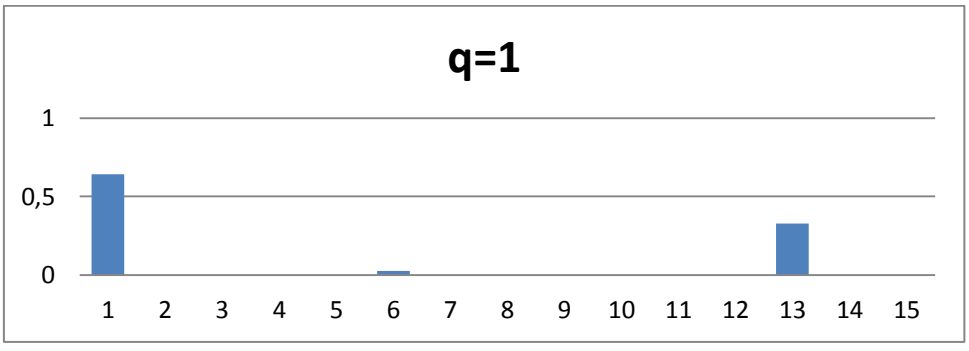
Στις καταστάσεις q=4, q=5, q=6 και q=7 έχουμε ακριβώς τα ίδια συμπεράσματα με τις προηγούμενες περιπτώσεις. Δηλαδή στην περίπτωση όπου q=4, q=6 και q=7 έχουμε συνεχώς νέες αναζητήσεις από τον χρήστη που σίγουρα υποδηλώνει μια πολύ σημαντική πνευματική κατάσταση την οποία μπορούμε να εκμεταλλευτούμε στην συνέχεια. Και η κατάσταση q=5 μας δείχνει την κατάσταση όπου η νέα αναζήτηση του είναι ή η τελευταία που κάνει κατά την συνεδρία του, είτε περνάει αρκετό χρόνο κοιτάζοντας τα αποτελέσματα

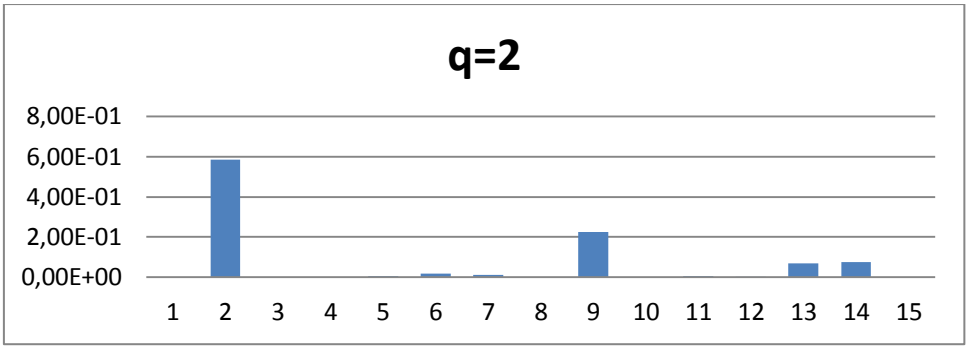
χωρίς όμως να αλληλεπιδράει καθόλου με την εφαρμογή, ούτε ακόμα μετακινώντας τον κέρσορα του.

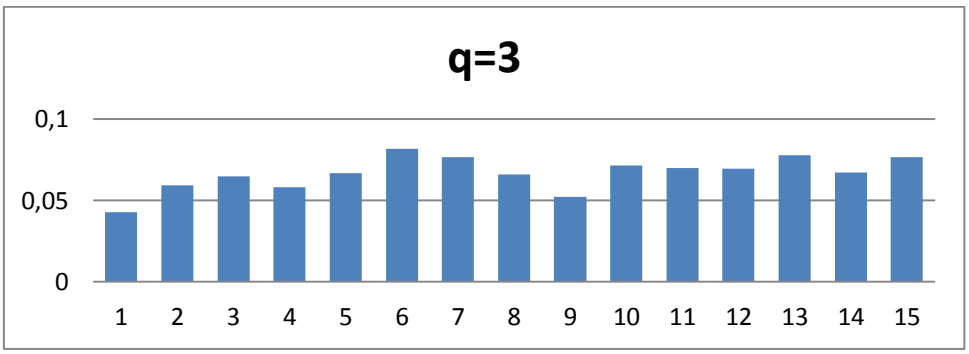
Με την παραπάνω ανάλυση που έχουμε επτά καταστάσεις στο μοντέλο μας, βλέπουμε πως σε αρκετές περιπτώσεις έχουμε επανάληψη των ίδιων καταστάσεων(overfit). Δηλαδή επαναλαμβάνεται η ίδια πνευματική κατάσταση του χρήστη ανάμεσα σε αυτά. Έτσι για λόγους πληρότητας κρίθηκε σκόπιμο να δούμε τα αντίστοιχα διαγράμματα και για την περίπτωση που έχουμε τέσσερις καταστάσεις. Αυτή ήταν και η δεύτερη επικρατέστερη περίπτωση και τελικά η πιο σωστή λόγω του προβλήματος που εμφανίστηκε με τις επτά καταστάσεις από την ανάλυση των πιθανοτήτων όπως περιγράψαμε στην παράγραφο 6.1 . Συγκεκριμένα, για τις περιπτώσεις των μοντέλων με αρχικές τιμές περίπου 1/ν , 1/μ καθώς και για τυχαίες αρχικές τιμές για να δούμε αν εμφανίζονται μονάχα οι καταστάσεις που περιγράψαμε και παραπάνω, χωρίς να υπάρχει επανάληψη τους.

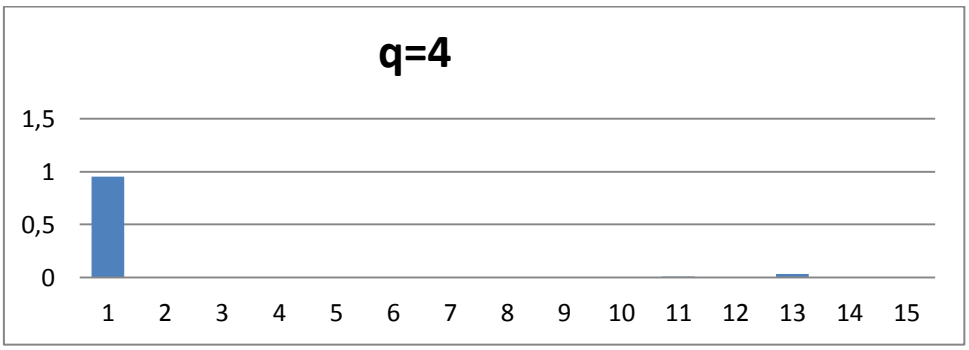
Έτσι έχοντας τέσσερις καταστάσεις, τα αποτελέσματα που προκύπτουν είναι τα παρακάτω για την περίπτωση που έχουμε μοντέλο με αρχικές τιμές περίπου 1/ν και 1/μ.

q/ id	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0,643256	6,05E-197	6,58E-101	5,41E-220	7,10E-45	0,027583	0,00E+00	4,25E-184	3,37E-61	4,64E-156	4,55E-212	1,02E-108	0,329161	2,41E-91	6,22E-53
2	7,45E-56	0,58595	4,89E-64	3,72E-93	0,005188	0,020095	0,012613	4,11E-162	0,226106	1,51E-23	0,004871	0,000265	0,068936	0,075977	8,26E-60
3	0,042734	0,05947	0,06466	0,058211	0,066574	0,081545	0,076733	0,065827	0,052292	0,071267	0,069744	0,06959	0,077629	0,067015	0,076708
4	0,952515	2,43E-62	0,004714	1,17E-169	5,59E-184	6,27E-143	2,54E-202	9,15E-209	6,95E-07	3,93E-147	0,009361	2,18E-108	0,033409	1,01E-215	2,69E-145









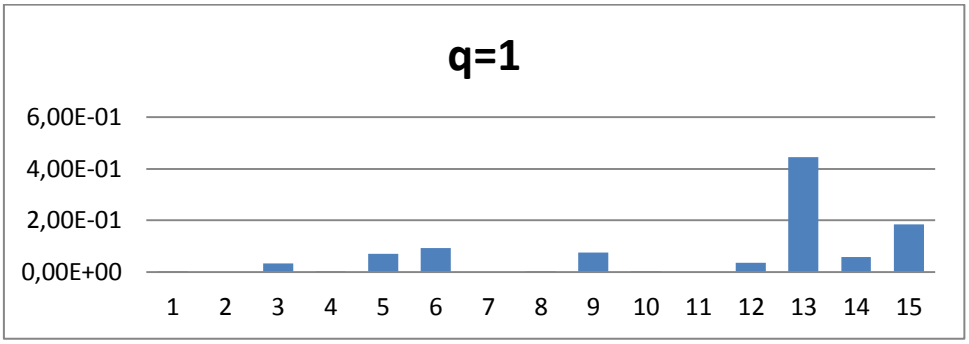
Από τις παραπάνω πνευματικές καταστάσεις των χρηστών παρατηρούμε πως κατά την εικόνα με q=1 έχουμε συσχέτιση των κινήσεων 1 και 13, εικόνα δηλαδή που αντιστοιχή στην κατάσταση q=5 που είδαμε κατά την ανάλυση των επτά καταστάσεων. Δηλαδή έχουμε τον χρήστη να κάνει συνεχώς νέες αναζητήσεις και μετα να παραμένει αδρανής, είτε διότι δεν τον ενδιέφεραν τελικά τα αποτελέσματα και εγκατέλειψε την εφαρμογή είτε γιατί περνάει αρκετό χρόνο κοιτάζοντας τα αποτελέσματα της αναζήτησης του.

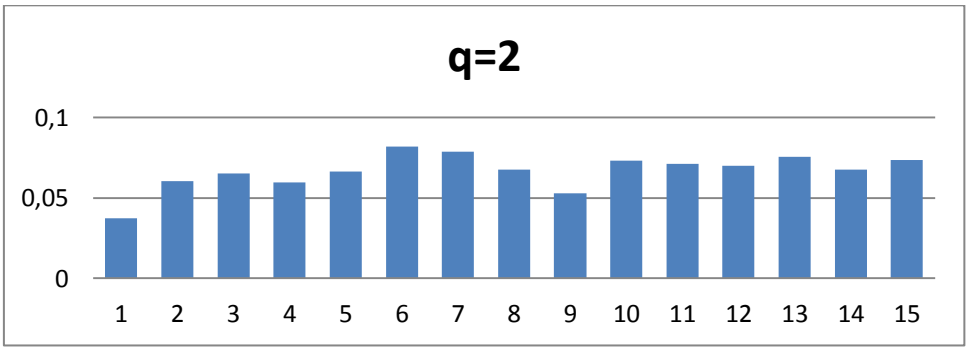
Στην συνέχεια κατά την πνευματική κατάσταση των χρηστών στην εικόνα με q=2 όπου έχουμε πολλές κινήσεις τύπου 2 και 9, και μερικές 13,14 έχουμε την αντίστοιχη κατάσταση με την εικόνα q=3 που αναλύσαμε κατά τις επτά καταστάσεις. Επίσης η εικόνα με q=3 όπου έχουμε ομαλή κατανομή των κινήσεων, έχουμε την περίπτωση q=2 κατά την ανάλυση του μοντέλου με τις επτά καταστάσεις, και τέλος κατά την πνευματική κατάσταση με q=4, όπου

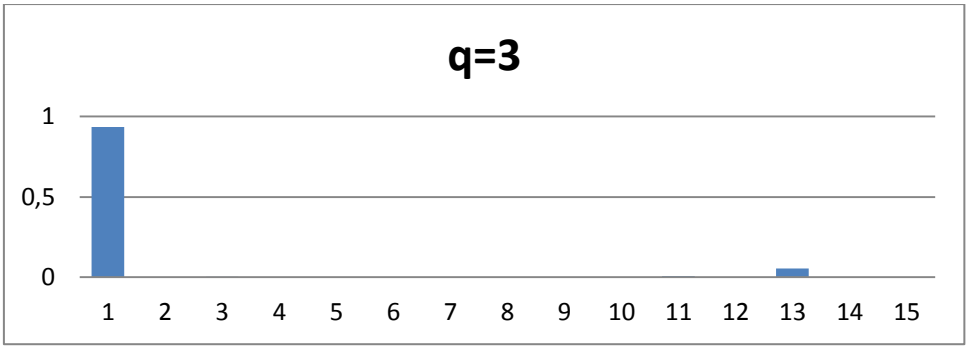
έχουμε πολλές νέες αναζητήσεις δηλαδή πολλές κινήσεις τύπου 1, και λίγες κινήσεις τύπου 13, έχουμε την αντίστοιχη περίπτωση των q=4, q=6 και q=7 που επαναλαμβάνονται. Βλέπουμε δηλαδή πως και με αυτό το μοντέλο των λιγότερων καταστάσεων έχουμε περιγράψει επιτυχώς όλες τις πνευματικές εικόνες εκτός από μία, και έχουμε αποφύγει τις περιπτώσεις των επαναλήψεων.

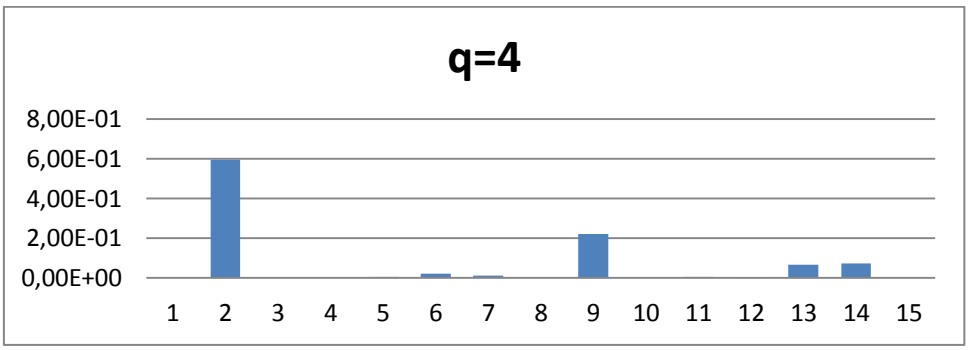
Και για την περίπτωση του μοντέλου με τυχαίες αρχικές τιμές τα αποτελέσματα που προκύπτουν είναι τα εξής:

q/ id	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	2,90E -05	1,88E -75	0,033 644	0,000 654	0,071 49	0,093 621	1,23E -52	9,01E -07	0,075 531	8,89E -07	2,91E -32	0,035 468	0,445 146	0,058 616	0,185 801
2	0,037 187	0,060 397	0,065 209	0,059 577	0,066 269	0,082 01	0,078 599	0,067 392	0,052 646	0,072 961	0,071 269	0,070 051	0,075 336	0,067 51	0,073 587
3	0,932 519	2,14E -35	0,004 287	2,12E -141	1,58E -141	4,43E -35	1,17E -183	1,95E -163	9,35E -09	9,45E -160	0,008 332	4,06E -144	0,054 862	1,96E -190	8,96E -119
4	1,52E -27	0,593 662	2,41E -30	2,77E -40	0,005 151	0,020 201	0,012 473	1,30E -67	0,221 525	1,45E -14	0,004 79	0,001 708	0,066 549	0,073 941	2,67E -32









Ομοίως και εδώ βρίσκουμε ομοιότητες σε σχέση με την προηγούμενη ανάλυση των επτά καταστάσεων που κάναμε. Συγκεκριμένα κατά την εικόνα με q=1 παρατηρούμε έχουμε πολλές τιμές των κινήσεων του χρήστη που κάνει συγκεκριμένη αναζήτηση και είναι η ίδια πνευματική κατάσταση με την q=1 κατά την προηγούμενη μας ανάλυση των επτά καταστάσεων. Επίσης , κατά την επόμενη εικόνα με q=2 έχουμε ομοιομορφία γύρω από όλες τις κινήσεις και είναι αντίστοιχη κατάσταση με την q=2 της προηγούμενης ανάλυσης. Στην συνέχεια παρατηρούμε πως η εικόνα με q=3 παρουσιάζει την κατάσταση με πολλές κινήσεις τύπου 1 και 13, και είναι αντίστοιχη με τις εικόνες με q=4, q=5 και q=6 της προηγούμενης ανάλυσης.

Παρομοίως και εδώ στο μοντέλο με τις τυχαίες αρχικές τιμές παρατηρούμε πως έχουμε απώλεια μιας πνευματικής κατάστασης του χρήστη, αλλά δεν έχουμε επανάληψη των καταστάσεων, άρα έχουμε ικανοποιητική περιγραφή και με αυτό το μοντέλο των τεσσάρων καταστάσεων.

Έτσι με χρήση των HMM βλέπουμε πως είναι δυνατόν να προκύψουν αυτές οι εικόνες των πνευματικών καταστάσεων των χρηστών και να μπορέσουμε να τις αποκωδικοποιήσουμε ώστε να ανακαλύψουμε πως προχωράει η σκέψη στο μυαλό του χρήστη. Αυτό μας επιτρέπει να βρούμε όλες τις δυνατές σκέψεις όπως αναλύσαμε παραπάνω και να μπορέσουμε να τις εκμεταλλευτούμε και ακόμη περισσότερο στην συνέχεια, καθώς με

βάση το μοντέλο που έχουμε περιγράψει των HMM και χρησιμοποιήσαμε για τις κινήσεις των χρηστών μπορούμε να κάνουμε αντίστοιχη εφαρμογή και στην πρόβλεψη των επόμενων καταστάσεων. Συγκεκριμένα, χρησιμοποιώντας το μοντέλο που έχουμε κατασκευάσει μπορούμε να υπολογίσουμε τις πιθανότητες για την επόμενη κίνηση του, τροφοδοτώντας με μια ακολουθία το μοντέλο μας και ελέγχοντας όλες τις πιθανές επόμενες κινήσεις να βρίσκουμε ποια επόμενη κίνηση έχει μεγαλύτερη πιθανότητα να προκύψει.

6.3 Παραδείγματα προβλέψεων των επομένων κινήσεων του χρήστη

Έχοντας κάνει πλεόν όλη την απαραίτητη ανάλυση για την ορθή και αποδοτική υλοποίηση του μοντέλου μας θα δώσουμε παραδείγματα για αυτή την περίπτωση όπου θέλουμε να υπολογίσουμε την πιθανότητα να προκύψει η επόμενη κίνηση του χρήστη. Έτσι χρησιμοποιώντας το μοντέλο των τεσσάρων κρυφών καταστάσεων και με αρχικές τιμές εκπαίδευσης περίπου 1/ν και 1/μ των πινάκων θα χρησιμοποιήσουμε τον αλγόριθμο Forward για πρόβλεψη επόμενης κίνησης, όπως θα δούμε και στα παρακάτω παραδείγματα.

Εάν μέχρι στιγμής οι κινήσεις του ήταν [1 1 13] . δηλαδή έκανε δύο αναζητήσεις και μετά έμεινε αδρανής. Έτσι εφαρμόζοντας τον αλγόριθμο Forward σύμφωνα με το όσα έχουμε αναλύσει παραπάνω για την υλοποίηση του και υπολογίζοντας τις πιθανότητες να προκύψουν οι ακολουθίες [1 1 13 x] όπου x=1,2,3,.....,15] βλέπουμε πως παίρνουμε τα παρακάτω αποτελέσματα όπου γίνεται εμφανές ότι μεγαλύτερη πιθανότητα για επόμενη κίνηση του χρήστη είναι η κίνηση 1, αμέσως μετά η κίνηση 2, και στη συνέχεια η κίνηση 3.

Ακολουθία [1 1 13]/ Επόμενη κίνηση	Πιθανότητα επόμενης κίνησης
1	0,0022
2	0,0016
3	0,00055
4	0,00048
5	0,00056
6	0,00073
7	0,00066

8	0,00055
9	0,00086
10	0,00059
11	0,00061
12	0,00058
13	0,0011
14	0,00071
15	0,00064

Πίνακας 2 – Πιθανότητες επόμενης κίνησης

Κάτι το οποίο είναι απολύτως ορθό κοιτάζοντας και τα διαγράμματα και την κατάσταση που μπορεί να βρίσκεται ο χρήστης, και να είναι ανάμεσα σε αυτές τις πνευματικές καταστάσεις όπως δείξαμε στο προηγούμενο κεφάλαιο 6.2.

Παρομοίως μπορούμε να υπολογίσουμε τις πιθανότητες και για οποιαδήποτε ακολουθία και να χρησιμοποιήσουμε έτσι το μοντέλο μας και για προβλέψεις , και για αυτό δίνουμε παρακάτω μερικά ακόμα παραδείγματα τα οποία αναλύονται και σχετίζονται όπως κάναμε και για την παραπάνω ακολουθία.

Επόμενη κίνηση	[1 2 9 2 2 2]	[1 2 2]	[1 2 9 9]	[1 6 10]	[2 9 13 14 2 9 9 2 2]
1	2.0440e-005	0,00044	0,000034	0,000054	1.3024e-008
2	4.2876e-004	0,0087	0,000650	0,000068	2.7310e-007
3	6.8852e-006	0,00018	0,000014	0,000071	4.3922e-009
4	6.1981e-006	0,00016	0,000012	0,000063	3.9539e-009
5	1.0829e-005	0,00026	0,000021	0,000072	6.9042e-009
6	2.3848e-005	0,00054	0,000041	0,000088	1.5198e-008
7	1.7264e-005	0,00039	0,000031	0,000083	1.1004e-008
8	7.0091e-006	0,00018	0,000014	0,000071	4.4712e-009
9	1.6857e-004	0,0035	0,000256	0,000058	1.0738e-007
10	7.5883e-006	0,00019	0,000015	0,000077	4.8407e-009
11	1.0938e-005	0,00026	0,000021	0,000075	6.9742e-009
12	7.6006e-006	0,00019	0,000015	0,000075	4.8483e-009
13	6.6062e-005	0,0014	0,000105	0,000086	4.2086e-008
14	6.1909e-005	0,0013	0,000098	0,000073	3.9439e-008
15	8.1676e-006	0,00021	0,000016	0,000083	5.2103e-009

Πίνακας 3 – Πιθανότητες επόμενων κινήσεων

Όπου κατά την ακολουθία [1 2 9 2 2 2] παρατηρούμε ότι η πιο πιθανή επόμενη κίνηση του χρήστη είναι η κίνηση 2 , η οποία έρχεται σε άμεση σχέση με το είδος αναζήτησης που κάνει και την πνευματική κατάσταση που έχουμε ήδη δει στο προηγούμενο κεφάλαιο, δηλαδή συνεχίζει την αναζήτηση του με άλλη μια ίδια κίνηση με τις προηγούμενες του.
Κατά την ακολουθία [1 2 2] προκύπτει ακριβώς η ίδια παρατήρηση με την προηγούμενη ακολουθία που αναλύσαμε δηλαδή με επόμενη κίνηση την 2.
Στην επόμενη ακολουθία [1 2 9 9] οι πιο πιθανές επόμενες κινήσεις είναι η κίνηση 2 και η κίνηση 9 αμέσως μετά, και κατά την ακολουθία [1 6 10] παρατηρούμε ότι όταν οι κινήσεις ενός χρήστη είναι τυχαίες όπως αυτή την περίπτωση και δεν μπορούν να περιγραφούν από κάποια πνευματική κατάσταση από τις παραπάνω, τότε είναι δύσκολο να προκύψει κάποια πρόβλεψη και όλες οι επόμενες κινήσεις παρουσιάζουν μικρές αποκλίσεις στις πιθανότητες να προκύψουν. Επίσης βασικός παράγοντας για την δυσκολία πρόβλεψης είναι και το μήκος της ακολουθίας που δεν είναι αρκετό για να προκύψει πρόβλεψη. Τέλος, κατά την ακολουθία [2 9 13 14 2 9 9 2 2] βλέπουμε πως υπάρχει αρκετά καλή πρόβλεψη για επόμενη κίνηση την κίνηση 2 καθώς και την κίνηση 9, κάτι που συμβαδίζει απόλυτα με την πνευματική κατάσταη των χρηστών που είδαμε και επίσης διότι έχουμε μια αρκετά μεγάλη ακολουθία 9 κινήσεων που μας επιτρέπει ακριβή πρόβλεψη.

Κεφάλαιο 7

Συμπεράσματα

7.1 Συμπεράσματα από τα πειράματα

Δείξαμε μέχρι στιγμής αναλυτικά όλο το προτεινόμενο μοντέλο με το οποίο μπορέσαμε να εκμεταλλευτούμε τα δεδομένα κατά τις αναζητήσεις των χρηστών και τις συνήθειες τους σε μια μηχανή αναζήτησης. Έτσι , εάν έχουμε στην διάθεση μας, μια μηχανή διερευνητικής αναζήτησης αυτές τις πληροφορίες των κινήσεων των χρηστών μπορούμε να τις αξιοποιήσουμε με τον ίδιο τρόπο για να δημιουργήσουμε το αντίστοιχο μοντέλο με αυτό. Φυσικά ανάλογα την εφαρμογή που μελετούμε θα διαφέρουν και τα πιθανά αποτελέσματα. Δείξαμε επίσης, πως πρέπει να διαχωρίσουμε τα κομμάτια της έρευνας μας, σε συνεδρίες, αναζητήσεις καθώς και τις αλληλεπιδράσεις των χρηστών, και πως να τα εκμεταλλευτούμε χωριστά. Τα αποτελέσματα που πήραμε επίσης από το μοντέλο μας, δείχνουν πως και στην πράξη κάτι τέτοιο είναι πολύ αποδοτικό για προβλέψεις των χρηστών καθώς και πόσο σημαντικές είναι για την σωστή κατασκευή του μοντέλου οι διάφορες προδιαγραφές που χρησιμοποιούμε, βασιζόμενοι στην θεωρία των Hidden Markov Models. Επίσης πρέπει να αναφέρουμε πως το μοντέλο μας μπορεί να γενικευτεί και για άλλες περιπτώσεις εκτός των μηχανών αναζήτησης. Πολύ σημαντικά κομμάτια στην επιτυχία του μοντέλου μας, είναι επίσης το πλήθος των δεδομένων που θα έχουμε στην διάθεση μας, το οποίο αφενώς πρέπει να είναι αρκετά μεγάλο ώστε να αντιπροσωπεύει το πλήθος των χρηστών, αλλά και όχι μεγαλύτερο από κάποιο μέγεθος καθώς θα μπορούσε κάτι τέτοιο αφενός να οδηγήσει σε λανθασμένη συμπεριφορά, αλλά και θα καθιστούσε την ανάπτυξη του πολύ χρονοβόρο και απαιτητικ;ο σε υπολογιστική ισχύ. Επίσης, με βάση τα μεγέθη που είχαμε και τα αποτελέσματα που πήραμε, βλέπουμε πως το μοντέλο μας που προκύπτει με τα HMM, περιγράφεται καλύτερα με την χρήση αρχικών τιμών εκπαίδευσης περίπου 1/ν και 1/μ , καθώς και πως ο βέλτιστος αριθμός καταστάσεων είναι οι επτά καταστάσεις καθώς μας δίνει την καλύτερη εικόνα σε σχέση με λιγότερες.

Παρατηρώντας την κατανομή των πιθανοτήτων στην συνέχεια στο μοντέλο HMM με τις επτά καταστάσεις που αναλύσαμε, βλέπουμε τις κατανομές που μας αποκαλύπτουν πως οι διάφορες κινήσεις των χρηστών ανάλογα την πνευματική κατάσταση που βρίσκονται

μαζεύονται γύρω από ορισμένες τιμές και προκύπτουν κάποια στοιχεία από αυτό. Συμπεράσματα που βγάζουμε όπως εξηγήσαμε και στο κομμάτι της αξιολόγησης κατά την διάρκεια ανάλυσης των αποτελεσμάτων αυτών είναι πχ ότι εάν μαζεύεται μεγάλος αριθμός τιμών από την κίνηση 13, που είναι η αδράνεια ενός χρήστη (idle) σημαίνει πως ο χρήστης λογικά έχει εγκαταλείψει την αναζήτηση του. Εάν πάλι έχουμε πολλές τιμές στο 1 που είναι η νέα αναζήτηση, ο χρήστης κάνει συνεχώς νέες αναζητήσεις που μας δείχνει στοιχεία για την κατάσταση που βρίσκεται δηλαδή πως δεν τον ικανοποιούν τα αποτελέσματα που παίρνει ή δεν είναι εξοικειωμένος με τις έξτρα λειτουργικότητες της εφαρμογής, ή ίσως απλά το θέμα που ψάχνει είναι πολύ άγνωστο και προσπαθεί να βρει από που θα αρχίσει την αναζήτηση του . Ομοίως και για τις υπόλοιπες τιμές βγαίνουν παρόμοια συμπεράσματα όπως εάν έχουμε πολλές τιμές απλωμένες σε όλο το φάσμα σημαίνει πως ο χρήστης ψάχνει γενικά όλες τις πληροφορίες, κάνοντας διερευνητική αναζήτηση, ή εάν έχουμε πολλές τιμές γύρω από τις τιμές 2,3 ο χρήστης ψάχνει συγκεκριμένους όρους κατά την αναζήτησης του. Επίσης, αυτά αλλάζουν ανάλογα την εικόνα δηλαδή από q=1 μέχρι q=7 , δηλαδή αλλάζουν και οι καταστάσεις που βρίσκονται οι χρήστες όπως περιγράψαμε παραπάνω, και υποδηλώνει διαφορετικούς τρόπους χρήσης της μηχανής αναζήτησης CRUISE. Και φυσικά όπως περιγράψαμε, η ουσία είναι η σύνδεση που συμβαίνει ανάμεσα σε αυτό το πλήθος κινήσεων σε μια συγκεκριμένη εικόνα και πως μπορούμε να την αποκωδικοποιήσουμε, καθώς και αν μπορούμε στην συνέχεια να εκμεταλλευτούμε περαιτέρω τα αποτελέσματα αυτά.

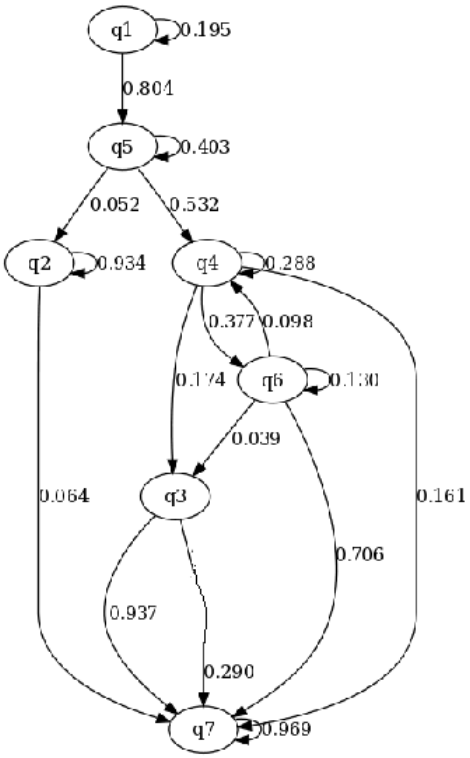
7.2 Επόμενα βήματα

Εφόσον πλέον έχουμε κατασκευάσει το μοντέλο όσον αφορά τις συνήθειες των χρηστών που μπορεί να προβλέπει επόμενες κινήσεις τους, θα πρέπει να ενσωματωθεί στην εφαρμογή CRUISE ώστε να γίνει ακόμα πιο εύχρηστη. Αυτό μπορεί να επιτευχθεί, χρησιμοποιώντας τις πιθανότητες από το μοντέλο μας, και εφόσον ο χρήστης έχει κάνει μια ακολουθία κινήσεων εμείς να του προτείνουμε τις πιο πιθανές επόμενες κινήσεις του, βοηθώντας τον καθ'όλη την διάρκεια της αναζήτησης του. Αυτό θα μπορούσε να συμβεί εάν η κάθε επόμενη κίνηση σημειωνόταν με πιο έντονο χρώμα ώστε να γίνει αντιληπτή από τον χρήστη αλλά και παράλληλα να είναι και διακριτική ώστε να μην ενοχλεί τους χρήστες και να τους παρασύρει σε άλλες επιλογές κατά την αναζήτηση τους. Αυτό μπορεί να γίνει

όπως είπαμε υπογραμμίζοντας επόμενες κινήσεις, είτε με κάποια μηνύματα αλληλεπίδρασης, που να το παροτρύνουν για νέα αναζήτηση αν είναι αρκετή ώρα στάσιμος κτλ.

Ακολουθώντας φυσικά την διαδικασία που περιγράψαμε στο κεφάλαιο 5, δηλαδή την δημιουργία του μοντέλου, μπορούμε να κάνουμε το πείραμα όλο και σε άλλες εφαρμογές όπως το CRUISE τα οποία κάνουν παρόμοιες λειτουργίες με αυτό, και να προκύψει και για αυτά αντίστοιχο μοντέλο. Στο δικό μας όμως κομμάτι, με το CRUISE θα μπορούσαμε στην συνέχεια να εκπαιδεύσουμε περισσότερο το μοντέλο , έχοντας περισσότερα στοιχεία από χρήστες με αποτέλεσμα να γίνουν οι προβλέψεις του ακόμα πιο ακριβείς.

Επίσης, όσον αφορά το κομμάτι που αναλύσαμε τελευταία με βάση τις καταστάσεις των χρηστών, έχοντας πλέον τις εικόνες των πνευματικών καταστάσεων τους όπως τις παρουσιάσαμε , υπάρχει και το επόμενο βήμα το οποίο θα είναι ο υπολογισμός των πιθανοτήτων να περνάει ο χρήστης από την μία κατάσταση στην άλλη, και είναι και αυτό το κομμάτι απαραίτητο ώστε να γνωρίζουμε ακόμα πιο εύκολα πιθανές επόμενες κινήσεις και να μπορεί η μηχανή να ανακαλύπτει ευκολότερα τις αλλαγές στην πνευματική κατάσταση που βρίσκεται ο χρήστης. Αυτή η ανάλυση μπορεί να συνεχιστεί με την χρήση του αλγορίθμου Viterbi [17], με τον οποίο μπορουμε να υπολογίσουμε τις πιθανότητες μετάβασης από κάθε μία πνευματική κατάσταση στην επόμενη. Δίνουμε και παρακάτω ένα ενδεικτικό παράδειγμα για το πως θα παρουσιάζονται οι πιθανότητες σε μια τέτοια περίπτωση.



Εικόνα 10 – Ενδεικτικές πιθανότητες υπολογισμένες με τον αλγόριθμο Viterbi(βασισμένο στο [17])

Δηλαδή, εφόσον κατασκευάσουμε το κομμάτι με τους πίνακες πιθανοτήτων στην συγκεκριμένη περίπτωση το HMM με τις 7 καταστάσεις, μετά παίρνουμε ακολουθίες και κάνουμε παρόμοια ανάλυση με αυτή που κάναμε στο προηγούμενο κομμάτι με την ανάλυση με τον αλγόριθμο Forward υπολογίζοντας το πιο πιθανό μονοπάτι. Σε αυτό το κομμάτι όμως δεν θα ασχοληθούμε με την υλοποίηση του στα πλαίσια της διπλωματικής αυτής, αλλά μπορεί να γίνει χρησιμοποιώντας τις πληροφορίες που έχουμε αναφέρει ήδη κατά την διάρκεια αυτής.

Τέλος να αναφέρουμε πως η ανάλυση και η μοντελοποίηση αυτή δεν είναι αποκλειστικά για το κομμάτι των μηχανών αναζήτησης μόνο, αλλά θα μπορούσε να τροποποιηθεί κατάλληλα ώστε να βρει εφαρμογή και σε άλλους τομείς που υπάρχουν τέτοιες λογικές ακολουθίες των χρηστών, βασιζόμενες στις προηγούμενες κινήσεις τους. Ένα χαρακτηριστικό παράδειγμα, θα μπορούσε να είναι η σειρά με την οποία ένα χρήστης ανοίγει τις εφαρμογές στο κινητό του (smartphone) τηλεφωνο [14]. Δηλαδή, αν έχει ανοίξει προηγουμένως κάποιες συγκεκριμένες εφαρμογές, μπορούμε με ανάλυση των κινήσεων αυτών να κάνουμε πρόβλεψη και για τις επόμενες που τον ενδιαφέρουν, και να συμβαίνει

αναδιάταξη των εφαρμογών στην επιφάνεια εργασίας του κινητού του με βάση την χρήση τους, αλλά και τις προβλέψεις που θα κάνουμε.

Κεφάλαιο 8

Παράρτημα

8.1 Κώδικας Matlab

8.1.1 Διάβασμα αρχείου Excel από το Matlab

Διάβασμα αρχείου Excel από το Matlab και μετατροπή του σε sequence μορφής cell array για σωστή είσοδο στην συνάρτηση hmmtrain που υλοποιεί τον αλγόριθμο Baum-Welch

```
filename=('C:\Users\.....\.....xlsx');
num=xlsread(filename,'C:C');
last=length(num);
e=unique(num);
n=length(e);
matr=[];
value=num(1);
counter=0;
for i=1:last
if(num(i)==value)
counter=counter+1;
else
matr=cat(1,matr,[counter]);
counter=1;
value=num(i);
end
end
matr=cat(1,matr,[counter]);
num=xlsread(filename,'D:D');
num=transpose(num);
matr=transpose(matr);
seq=mat2cell(num,[1],matr);
```

Στην στήλη C του αρχείου Excel είναι τα session id, και στην στήλη D τα actions. Στο τέλος αποθηκεύεται στην μεταβλητή seq η ακολουθία που περιέχει όλα τα sessions.

8.1.2 Μοντελοποίηση και αξιολόγηση του μοντέλου

Στον παρακάτω κώδικα έχουμε την μοντελοποίηση με την χρήση του αλγορίθμου Baum-Welch και την αξιολόγηση με την χρήση του αλγορίθμου Forward με βάση την διαδικασία που περιγράφεται στο κεφάλαιο 6.

Με τιμές περίπου 1/n , 1/m έχουμε :

A2 = [0.45,0.55;0.53,0.47];

B2 = [0.06, 0.06, 0.06, 0.06, 0.07, 0.06, 0.07, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06;0.06, 0.06, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.07, 0.06, 0.06, 0.06, 0.07];

A3 = [0.35,0.35,0.3;0.33,0.31,0.36;0.3,0.34,0.36];

B3 = [0.07, 0.06, 0.06, 0.06, 0.06, 0.07, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06;0.06, 0.06, 0.06, 0.06, 0.07, 0.06, 0.07, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06;0.06, 0.06, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.07, 0.06, 0.06, 0.06, 0.07];

A4 = [0.25,0.22,0.25,0.28;0.24,0.23,0.26,0.27;0.25,0.25,0.25,0.25;0.21,0.22,0.28,0.29];

B4 = [0.07, 0.07, 0.06, 0.06, 0.06, 0.06, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06;0.06, 0.06, 0.06, 0.06, 0.07, 0.06, 0.07, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06;0.06, 0.06, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.07, 0.06, 0.06, 0.06, 0.07;0.07, 0.06, 0.06, 0.06, 0.06, 0.07, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06];

A5 = [0.18,0.22,0.23,0.17,0.19;0.21,0.2,0.21,0.19,0.2;0.18,0.22,0.23,0.17,0.19;0.18,0.22,0.22,0.18,0.19;0.21,0.2,0.21,0.19,0.2];

B5 = [0.06, 0.07, 0.06, 0.06, 0.06, 0.07, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06;0.06, 0.06, 0.06, 0.06, 0.07, 0.06, 0.07, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06;0.06, 0.06, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.07, 0.06, 0.06, 0.06, 0.07;0.07, 0.06, 0.06, 0.06, 0.06, 0.07, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06;0.07, 0.06, 0.06, 0.06, 0.06, 0.07, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06];

A6 = [0.16,0.17,0.16,0.17,0.17,0.17;0.16,0.17,0.16,0.18,0.16,0.17;0.18,0.15,0.16,0.17,0.17,0.17;0.16,0.17,0.16,0.17,0.17,0.17;0.18,0.15,0.16,0.17,0.17,0.17;0.18,0.15,0.16,0.17,0.17,0.17];

B6 = [0.06, 0.07, 0.06, 0.06, 0.06, 0.07, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06;0.06, 0.06, 0.06, 0.06, 0.07, 0.06, 0.07, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06;0.06, 0.06, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.07, 0.06, 0.06, 0.06, 0.07;0.07, 0.06, 0.06, 0.06, 0.06, 0.07, 0.06, 0.06,

0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06;0.07, 0.06, 0.06, 0.06, 0.06, 0.07, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06;0.07, 0.06, 0.06, 0.06, 0.06, 0.07, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06];

A7 = [0.14,0.14,0.13,0.15,0.14,0.15,0.15;0.14,0.14,0.13,0.15,0.14,0.15,0.15;0.14,0.14,0.13,0.15,0.14,0.15,0.15;0.14,0.14,0.13,0.15,0.14,0.15,0.15;0.14,0.14,0.13,0.15,0.14,0.15,0.15;0.14,0.14,0.13,0.15,0.14,0.15,0.15;0.14,0.14,0.13,0.15,0.14,0.15,0.15];

B7 = [0.06, 0.07, 0.06, 0.06, 0.06, 0.07, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06;0.06, 0.06, 0.06, 0.06, 0.07, 0.06, 0.07, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06;0.06, 0.06, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.07, 0.06, 0.06, 0.06, 0.07;0.07, 0.06, 0.06, 0.06, 0.06, 0.07, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06;0.07, 0.06, 0.06, 0.06, 0.06, 0.07, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06;0.07, 0.06, 0.06, 0.06, 0.06, 0.07, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06];

Με τιμές ακριβώς 1/n , 1/m έχουμε :

A2 = [0.5,0.5;0.5,0.5];

B2 = [0.066, 0.066, 0.066, 0.066, 0.066, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067;0.066, 0.066, 0.066, 0.066, 0.066, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067];

A3 = [0.33,0.33,0.34;0.33,0.34,0.33;0.34,0.33,0.33];

B3 = [0.066, 0.066, 0.066, 0.066, 0.066, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067;0.066, 0.066, 0.066, 0.066, 0.066, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067;0.066, 0.066, 0.066, 0.066, 0.066, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067];

A4 = [0.25,0.25,0.25,0.25;0.25,0.25,0.25,0.25;0.25,0.25,0.25,0.25;0.25,0.25,0.25,0.25];

B4 = [0.066, 0.066, 0.066, 0.066, 0.066, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067;0.066, 0.066, 0.066, 0.066, 0.066, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067;0.066, 0.066, 0.066, 0.066, 0.066, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067;0.066, 0.066, 0.066, 0.066, 0.066, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067];

A5 = [0.2,0.2,0.2,0.2,0.2;0.2,0.2,0.2,0.2,0.2;0.2,0.2,0.2,0.2,0.2;0.2,0.2,0.2,0.2,0.2;0.2,0.2,0.2,0.2,0.2];

B5 = [0.066, 0.066, 0.066, 0.066, 0.066, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067;0.066, 0.066, 0.066, 0.066, 0.066, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067;0.066, 0.066, 0.066, 0.066, 0.066, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067;0.066, 0.066, 0.066, 0.066, 0.066, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067;0.066, 0.066, 0.066, 0.066, 0.066, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067];

A6 = [0.166,0.166,0.167,0.167,0.167,0.167;0.166,0.166,0.167,0.167,0.167,0.167;0.166,0.166,0.167,0.167,0.167,0.167;0.166,0.166,0.167,0.167,0.167,0.167;0.166,0.166,0.167,0.167,0.167,0.167;0.166,0.166,0.167,0.167,0.167,0.167];

B6 = [0.066, 0.066, 0.066, 0.066, 0.066, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067;0.066, 0.066, 0.066, 0.066, 0.066, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067;0.066, 0.066, 0.066, 0.066, 0.066, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067;0.066, 0.066, 0.066, 0.066, 0.066, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067;0.066, 0.066, 0.066, 0.066, 0.066, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067;0.066, 0.066, 0.066, 0.066, 0.066, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067, 0.067];

A7 = [0.143,0.143,0.143,0.142,0.143,0.143,0.143;0.143,0.143,0.143,0.142,0.143,0.143,0.143;0.143,0.143,0.143,0.142,0.143,0.143,0.143;0.143,0.143,0.143,0.142,0.143,0.143,0.143;0.143,0.143,0.143,0.142,0.143,0.143,0.143;0.143,0.143,0.143,0.142,0.143,0.143,0.143;0.143,0.143,0.143,0.142,0.143,0.143,0.143];

B7 = [0.06, 0.07, 0.06, 0.06, 0.06, 0.07, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06;0.06, 0.06, 0.06, 0.06, 0.07, 0.06, 0.07, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06;0.06, 0.06, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.07, 0.06, 0.06, 0.06, 0.07;0.07, 0.06, 0.06, 0.06, 0.06, 0.07, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06;0.07, 0.06, 0.06, 0.06, 0.06, 0.07, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06;0.07, 0.06, 0.06, 0.06, 0.06, 0.07, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06;0.07, 0.06, 0.06, 0.06, 0.06, 0.07, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06];

Με τυχαίες αρχικές τιμές έχουμε :

A2 = [0.8,0.2;0.3,0.7];

B2 = [0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.05, 0.05, 0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.1;0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.05, 0.05, 0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.1];

A3 = [0.1,0.6,0.3;0.8,0.1,0.1;0.2,0.3,0.5];

B3 = [0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.05, 0.05, 0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.1;0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.05, 0.05, 0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.1;0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.05, 0.05, 0.1, 0.2, 0.05, 0.01, 0.03, 0.1, 0.01];

A4 = [0.2,0.3,0.2,0.3;0.6,0.1,0.1,0.2;0.2,0.3,0.2,0.3;0.2,0.3,0.2,0.3];

B4 = [0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.05, 0.05, 0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.1;0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.05, 0.05, 0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.1;0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.05, 0.05, 0.1, 0.2, 0.05, 0.01, 0.03, 0.1, 0.01;0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.05, 0.05, 0.1, 0.2, 0.05, 0.01, 0.03, 0.1, 0.01];

A5 = [0.1,0.1,0.1,0.2,0.5;0.5,0.1,0.1,0.1,0.2;0.1,0.1,0.1,0.1,0.6;0.6,0.1,0.1,0.1,0.1;0.5,0.1,0.1,0.1,0.2];

B5 = [0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.05, 0.05, 0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.1;0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.05, 0.05, 0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.1;0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.05, 0.05, 0.1, 0.2, 0.05, 0.01, 0.03, 0.1, 0.01;0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.05, 0.05, 0.1, 0.2, 0.05, 0.01, 0.03, 0.1, 0.01;0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.05, 0.05, 0.1, 0.2, 0.05, 0.01, 0.03, 0.1, 0.01];

A6 = [0.05,0.1,0.15,0.3,0.1,0.3;0.05,0.1,0.15,0.3,0.3,0.1;0.1,0.05,0.15,0.3,0.1,0.3;0.05,0.1,0.15,0.3,0.1,0.3;0.05,0.1,0.15,0.3,0.1,0.3;0.05,0.1,0.15,0.3,0.1,0.3];

B6 = [0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.05, 0.05, 0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.1;0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.05, 0.05, 0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.1;0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.05, 0.05, 0.1, 0.2, 0.05, 0.01, 0.03, 0.1, 0.01;0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.05, 0.05, 0.1, 0.2, 0.05, 0.01, 0.03, 0.1, 0.01;0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.05, 0.05, 0.1, 0.2, 0.05, 0.01, 0.03, 0.1, 0.01;0.1, 0.05, 0.2, 0.01, 0.03, 0.01, 0.05, 0.05, 0.1, 0.2, 0.05, 0.01, 0.03, 0.1, 0.01];

A7 = [0.1,0.05,0.05,0.2,0.1,0.2,0.3;0.1,0.05,0.05,0.2,0.1,0.2,0.3;0.1,0.05,0.05,0.2,0.1,0.2,0.3;0.1,0.05,0.05,0.2,0.1,0.2,0.3;0.1,0.05,0.05,0.2,0.1,0.2,0.3;0.1,0.05,0.05,0.2,0.1,0.2,0.3];

B7 = [0.06, 0.07, 0.06, 0.06, 0.06, 0.07, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06;0.06, 0.06, 0.06, 0.06, 0.07, 0.06, 0.07, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06;0.06, 0.06, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.07, 0.06, 0.06, 0.06, 0.07;0.07, 0.06, 0.06, 0.06, 0.06, 0.07, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06;0.07, 0.06, 0.06, 0.06, 0.06, 0.07, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06;0.07, 0.06, 0.06, 0.06, 0.06, 0.07, 0.06, 0.06, 0.07, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06];

Και οι τιμές για την έναρξη του Forward είναι οι ίδιες σε κάθε περίπτωση που έχουμε διαφορετικές αρχικές τιμές απλά διαφέρουν ανάλογα με τον αριθμό των hidden states και είναι οι παρακάτω :

```
pi2=[0.5 0.5];

pi3=[0.33 0.33 0.34];

pi4=[0.25 0.25 0.25 0.25];

pi5=[0.2 0.2 0.2 0.2 0.2];

pi6=[0.166 0.166 0.167 0.167 0.167 0.167];

pi7=[0.143 0.143 0.143 0.142 0.143 0.143 0.143];
```

Πέρα από τις αρχικές τιμές , ακολουθεί και ο κώδικας που χρησιμοποιήθηκε στο Matlab για τον υπολογισμό όσων έχουμε περιγράψει κατά την διάρκεια της εργασίας στο κεφάλαιο 6. Δίνεται ο κώδικας για την περίπτωση με αρχικές τιμές περίπου 1/n , 1/m, για τον κώδικα των υπολοίπων περιπτώσεων απλά αλλάζουμε τις τιμές των μεταβλητών για να πάρουμε τους άλλους πίνακες A και B .

```
for i=1:n

news=seq;

news(i)=[];

seque1=seq(i);

seque1=cell2mat(seque1);

[A2Hat, B2Hat] = hmmtrain(news,A2,B2);

A12{i}=A2Hat;

B12{i}=B2Hat;

p12(i)=pr_hmm(seque1,A2Hat,B2Hat,pi2);

for j=1:n

seque2=seq(j);

seque2=cell2mat(seque2);
```

```
p212(i,j)=pr_hmm(seque2,A2Hat,B2Hat,pi2);

end

end

for i=1:n

news=seq;

news(i)=[];

seque1=seq(i);

seque1=cell2mat(seque1);

[A3Hat, B3Hat] = hmmtrain(news,A3,B3);

A13{i}=A3Hat;

B13{i}=B3Hat;

p13(i)=pr_hmm(seque1,A3Hat,B3Hat,pi3);

for j=1:n

seque2=seq(j);

seque2=cell2mat(seque2);

p213(i,j)=pr_hmm(seque2,A3Hat,B3Hat,pi3);

end

end

for i=1:n

news=seq;

news(i)=[];

seque1=seq(i);

seque1=cell2mat(seque1);

[A4Hat, B4Hat] = hmmtrain(news,A4,B4);

A14{i}=A4Hat;
```

```
B14{i}=B4Hat;

p14(i)=pr_hmm(seque1,A4Hat,B4Hat,pi4);

for j=1:n

seque2=seq(j);

seque2=cell2mat(seque2);

p214(i,j)=pr_hmm(seque2,A4Hat,B4Hat,pi4);

end

end



for i=1:n

news=seq;

news(i)=[];

seque1=seq(i);

seque1=cell2mat(seque1);

[A5Hat, B5Hat] = hmmtrain(news,A5,B5);

A15{i}=A5Hat;

B15{i}=B5Hat;

p15(i)=pr_hmm(seque1,A5Hat,B5Hat,pi5);

for j=1:n

seque2=seq(j);

seque2=cell2mat(seque2);

p215(i,j)=pr_hmm(seque2,A5Hat,B5Hat,pi5);

end

end
```

```
for i=1:n
news=seq;
news(i)=[];
seque1=seq(i);
seque1=cell2mat(seque1);
[A6Hat, B6Hat] = hmmtrain(news,A6,B6);
A16{i}=A6Hat;
B16{i}=B6Hat;
p16(i)=pr_hmm(seque1,A6Hat,B6Hat,pi6);
for j=1:n
seque2=seq(j);
seque2=cell2mat(seque2);
p216(i,j)=pr_hmm(seque2,A6Hat,B6Hat,pi6);
end
end

for i=1:n
news=seq;
news(i)=[];
seque1=seq(i);
seque1=cell2mat(seque1);
[A7Hat, B7Hat] = hmmtrain(news,A7,B7);
A17{i}=A7Hat;
B17{i}=B7Hat;
p17(i)=pr_hmm(seque1,A7Hat,B7Hat,pi7);
```

```
for j=1:n

seque2=seq(j);

seque2=cell2mat(seque2);

p217(i,j)=pr_hmm(seque2,A7Hat,B7Hat,pi7);

end

end
```

Στις μεταβλητές (πίνακες θέσεων 231x231) A12,A13,.....,A17 και B12,13,....B17 αποθηκεύονται τα επιμέρους μοντέλα, και στις μεταβλητές (πίνακες 1x231) p12,p13,….,p17 αποθηκεύονται τα αποτελέσματα της πρώτης μεθόδου που περιγράψαμε για αξιολόγηση των μοντέλων, και τέλος στις μεταβλητές (πίνακες 231x231) p212,p213,….p217 τα αποτελέσματα της δεύτερης μεθόδου που περιγράψαμε.

Τέλος για να πάρουμε τους λογαρίθμους, τις μέσες τιμές και τα αθροίσματα, χρησιμοποιούμε τις εντολές log10(); mean(); sum() καταλλήλως.

Και δίνεται στην συνέχεια και μια υλοποίηση του αλγορίθμου forward που χρησιμοποιήσαμε:

```
function p=pr_hmm(o,a,b,pi)
%INPUTS:
%O=Given observation sequence labellebd in numerics
%A(N,N)=transition probability matrix
%B(N,M)=Emission matrix
%pi=initial probability matrix
%Output
%P=probability of given sequence in the given model
n=length(a(1,:));
T=length(o);
%it uses forward algorith to compute the probability
for i=1:n       %it is initilization
   m(1,i)=b(i,o(1))*pi(i);
end
for t=1:(T-1)     %recurssion
   for j=1:n
```

```
      z=0;
      for i=1:n
         z=z+a(i,j)*m(t,i);
      end
      m(t+1,j)=z*b(j,o(t+1));
   end
end
p=0;
for i=1:n        %termination
   p=p+m(T,i);
end
```

8.2 Κώδικας JQuery,Javascript

8.2.1 Νέα Αναζήτηση

```
$.ajax({
        url:baseUrl+'/terms?q='+searchTerms+/*'&level='+level+*/'&source='+source
      }).done(function(data){
        $("#cloud").html(data);
        $("#cloud-wrapper").fadeIn('fast');
        //$("#query-wrapper").fadeIn('fast');
       //Get Recommendations from Social Recommender
        getRecommendations(searchTerms);
        var now1 = new Date();
        var now1_utc = new Date(now1.getUTCFullYear(), now1.getUTCMonth(), now1.getUTCDate(),
now1.getUTCHours(), now1.getUTCMinutes(), now1.getUTCSeconds());
        var time1=now1_utc.toString();
        var typo1="New Query|"
        var typotostr=ivalue.toString();
        var typo2=typo1.concat(typotostr);
        var typo3=typo2.concat("|");
        var typo4=typo3.concat(searchTerms);
        var typo5=typo4.concat("|");
        var typo6=typo5.concat(time1);
        actionsarray.push(typo6);
          shareCon4(typo6);
        }
```

```
        ivalue=[];
        iccs.handleStep(2);
```

8.2.2 Ταχύτητα κέρσορα

```
var mrefreshinterval = 500; // update display every 500ms
 var lastmousex=-1;
 var lastmousey=-1;
 var lastmousetime;
 //var mousetravel = 0;
 $('html').mousemove(function(e) {
    var mousex = e.pageX;
    var mousey = e.pageY;
    if (lastmousex > -1) {
       mousetravel += Math.max( Math.abs(mousex-lastmousex), Math.abs(mousey-lastmousey) );
       mousemovearray.push(mousetravel);
       mousetravel=0;
    }
    lastmousex = mousex;
    lastmousey = mousey;
 });
```

8.2.3 Hover over image

```
        var now1 = new Date();
        var now1_utc = new Date(now1.getUTCFullYear(), now1.getUTCMonth(), now1.getUTCDate(),
now1.getUTCHours(),        now1.getUTCMinutes(), now1.getUTCSeconds());
        var time1=now1_utc.toString();
        var typo1="Hover|";
        var typo2=typo1.concat("Image|");
        var typo3=typo2.concat($title);
        var typo4=typo3.concat("|");
        var typotostr2=$url;
        typotostr2=typotostr2.toString();
        var typo5=typo4.concat(typotostr2);
        var typo6=typo5.concat("|");
        var typo7=typo6.concat($thumb);
```

```
        var typo8=typo7.concat("|");
        var typo9=typo8.concat(time1);
        actionsarray.push(typo9);
          shareCon4(typo9);
                actionscounter=0;
                actionsarray=[];
        }
```

8.2.4 Click on image

```
$cloud.click(function(){
          if($(this).find("img").attr("alt")==="Image"){//} length>0){        //has('img')
             var $thumb= $(this).find("img").attr("src");
             var $url = $(this).find(".image-class").attr("href");
          var $title = $(this).attr("data-title");
        var now1 = new Date();
        var now1_utc = new Date(now1.getUTCFullYear(), now1.getUTCMonth(), now1.getUTCDate(),
now1.getUTCHours(), now1.getUTCMinutes(), now1.getUTCSeconds());
        var time1=now1_utc.toString();
        var typo1="Click|";
        var typo2=typo1.concat("Image|");
        var typo3=typo2.concat($title);
        var typo4=typo3.concat("|");
        var typo5=typo4.concat($url);
        var typo6=typo5.concat("|");
        var typo7=typo6.concat($thumb);
        var typo8=typo7.concat("|");
        var typo9=typo8.concat(time1);
        actionsarray.push(typo9);
          shareCon4(typo9);
        actionscounter++;
        if(actionscounter>=10){
                for (i = 0; i < (actionscounter); i++) {
                  //shareCon3(actionsarray[i]);//<<--write to database
                 }
                actionscounter=0;
                actionsarray=[];

        }
```

8.2.5 Add image to bucket

```
$('.btn-cloud').bind('click',function(e){
        e.preventDefault();
        var $action = $(this).attr("data-action");
        if($action =='boost'){
          if($(this).parent().parent().find("img").attr("alt")==="Image"){//} length>0){
            var $thumb= $(this).parent().parent().find("img").attr("src");
             var $url = $(this).parent().parent().find(".image-class").attr("href");
          var $title = $(this).parent().parent().attr("data-title");
            var $semanticgroup = $(this).parent().parent().attr("data-semanticgroup");
        var now1 = new Date();
        var now1_utc = new Date(now1.getUTCFullYear(), now1.getUTCMonth(), now1.getUTCDate(),
now1.getUTCHours(), now1.getUTCMinutes(), now1.getUTCSeconds());
        var time1=now1_utc.toString();
        var typo1="Click bucket|";
        var typo2=typo1.concat("Image|");
        var typo3=typo2.concat($title);
        var typo4=typo3.concat("|");
        var typo5=typo4.concat($url);
        var typo6=typo5.concat("|");
        var typo7=typo6.concat($thumb);
        var typo8=typo7.concat("|");
        var typo9=typo8.concat(time1);
        actionsarray.push(typo9);
          shareCon4(typo9);
        actionscounter++;
          $.publish("image.added",[$thumb,$url,$title,$semanticgroup]);
          }
```

8.2.6 Click on text for extra info

```
$('.btn-tweets').clickover({
        animation:true,
        html:true,
        placement:function(tip,element){
          var $cloud = $("#wordcloud");
```

```
            var offset = $(element).parent().parent().position();
            var popupWidth=300;
            var popupHeight=190;
            var height = $cloud.innerHeight();
            var width = $cloud.innerWidth();
            console.log("Dimension %sx%s",width,height);
            console.log("Element offset",offset);
            var vert="bottom";
            if(offset.top+ popupHeight >= height){
               vert = "top";
            }
            var hor="Right";
            if(offset.left+ popupWidth >= width){
               hor="Left";
            }
            return vert+hor;
         },
         'content':function(){
            $word = $(this).parent().parent();
            var id =$word.attr("data-id");
            var source=$word.attr("data-source");
            var $term3= $(this).parent().parent().text().trim();
         var now1 = new Date();
         var now1_utc = new Date(now1.getUTCFullYear(), now1.getUTCMonth(), now1.getUTCDate(),
now1.getUTCHours(), now1.getUTCMinutes(), now1.getUTCSeconds());
         var time1=now1_utc.toString();
         var typo1="Click fakos|";
         var typo2=typo1.concat("Text|");
         var typo3=typo2.concat($term3);
         var typo4=typo3.concat("|");
         var typo5=typo4.concat(time1);
         actionsarray.push(typo5);
            shareCon4(typo5);
         actionscounter++;
            var arr = $term3.split(", ");
            var unique = [];
            $.each(arr, function (index,word) {
              if ($.inArray(word, unique) === -1)
```

```
            unique.push(word);
        });
        var $term4=unique.toString();
        var $ret = $('<div class="popover-tweet-content"/>');
        $ret.html($("#term-tweets-"+id).html());
        return $ret;
    }
});
```

8.3 Δείγμα αρχείου βάσης δεδομένων

Ακολουθεί ένα δείγμα από το αρχείο των κινήσεων των χρηστών (Log) που χρησιμοποιήσαμε για την εκπαίδευση του μοντέλου. Συγκεκριμένα η πρώτη στήλη Id περιλαμβάνει τον μοναδικό αριθμό της κάθε καταχώρησης στην βάση δεδομένων , η επόμενη στήλη User είναι το όνομα του χρήστη, το Session Id είναι ο μοναδικός αριθμός της κάθε συνεδρίας, Action Id είναι ο κωδικός της κάθε κίνησης του χρήστη, κωδικοποιημένη με βάση όσα αναλύσαμε στο κεφάλαιο 4. Και τέλος Date είναι η χρονική στιγμή που έγινε η κάθε κίνηση. Δίνουμε παρακάτω τα αποτελέσματα δύο συνεδριών από δύο διαφορετικούς χρήστες.

Id	User	Session Id	Action Id	Date
840	kostis	421	1	Wed Aug 27 2014 15:55:51 GMT+0300 (GTB Daylight Time)
846	kostis	421	2	Wed Aug 27 2014 15:55:51 GMT+0300 (GTB Daylight Time)
847	kostis	421	2	Wed Aug 27 2014 15:55:56 GMT+0300 (GTB Daylight Time)
850	kostis	421	2	Wed Aug 27 2014 15:55:56 GMT+0300 (GTB Daylight Time)
851	kostis	421	2	Wed Aug 27 2014 15:56:04 GMT+0300 (GTB Daylight Time)
854	kostis	421	2	Wed Aug 27 2014 15:56:04 GMT+0300 (GTB Daylight Time)
855	kostis	421	2	Wed Aug 27 2014 15:56:31 GMT+0300 (GTB Daylight Time)
883	kostis	421	9	Wed Aug 27 2014 15:56:33 GMT+0300 (GTB Daylight Time)
888	kostis	421	14	Wed Aug 27 2014 15:57:07 GMT+0300 (GTB Daylight Time)

889	kostis	421	1	Wed Aug 27 2014 15:57:37 GMT+0300 (GTB Daylight Time)
903	kostis	421	1	Wed Aug 27 2014 15:57:41 GMT+0300 (GTB Daylight Time)
906	kostis	421	6	Wed Aug 27 2014 15:57:55 GMT+0300 (GTB Daylight Time)
913	kostis	421	6	Wed Aug 27 2014 15:58:01 GMT+0300 (GTB Daylight Time)
915	kostis	421	6	Wed Aug 27 2014 15:58:04 GMT+0300 (GTB Daylight Time)
918	kostis	421	6	Wed Aug 27 2014 15:58:04 GMT+0300 (GTB Daylight Time)
917	kostis	421	10	Wed Aug 27 2014 15:58:06 GMT+0300 (GTB Daylight Time)
920	kostis	421	14	Wed Aug 27 2014 15:58:06 GMT+0300 (GTB Daylight Time)
921	kostis	421	13	Wed Aug 27 2014 16:05:59 GMT+0300 (GTB Daylight Time)
922	kostis	421	1	Wed Aug 27 2014 16:06:06 GMT+0300 (GTB Daylight Time)
933	kostis	421	6	Wed Aug 27 2014 16:06:10 GMT+0300 (GTB Daylight Time)
935	kostis	421	6	Wed Aug 27 2014 16:06:14 GMT+0300 (GTB Daylight Time)
938	kostis	421	6	Wed Aug 27 2014 16:06:14 GMT+0300 (GTB Daylight Time)
937	kostis	421	10	Thu Aug 21 2014 11:49:05 GMT+0300 (GTB Daylight Time)
1	martar	312	1	Thu Aug 21 2014 11:49:17 GMT+0300 (GTB Daylight Time)
7	martar	312	9	Thu Aug 21 2014 11:49:20 GMT+0300 (GTB Daylight Time)
8	martar	312	2	Thu Aug 21 2014 11:49:20 GMT+0300 (GTB Daylight Time)
9	martar	312	2	Thu Aug 21 2014 11:49:24 GMT+0300 (GTB Daylight Time)
11	martar	312	2	Thu Aug 21 2014 11:49:24 GMT+0300 (GTB Daylight Time)
12	martar	312	2	Thu Aug 21 2014 11:49:26 GMT+0300 (GTB Daylight Time)
13	martar	312	7	Thu Aug 21 2014 11:49:40 GMT+0300 (GTB Daylight Time)
17	martar	312	2	Thu Aug 21 2014 11:49:40 GMT+0300 (GTB Daylight Time)
16	martar	312	2	Thu Aug 21 2014 11:49:43 GMT+0300 (GTB Daylight Time)

18	martar	312	9	Thu Aug 21 2014 11:49:46 GMT+0300 (GTB Daylight Time)
20	martar	312	14	Thu Aug 21 2014 11:51:50 GMT+0300 (GTB Daylight Time)
59	martar	312	15	Thu Aug 21 2014 11:52:04 GMT+0300 (GTB Daylight Time)
121	martar	312	13	Thu Aug 21 2014 11:52:14 GMT+0300 (GTB Daylight Time)
60	martar	312	1	Thu Aug 21 2014 11:54:05 GMT+0300 (GTB Daylight Time)
128	martar	312	15	Thu Aug 21 2014 11:54:14 GMT+0300 (GTB Daylight Time)
151	martar	312	15	Thu Aug 21 2014 11:54:50 GMT+0300 (GTB Daylight Time)
161	martar	312	2	Thu Aug 21 2014 11:54:50 GMT+0300 (GTB Daylight Time)
160	martar	312	2	Thu Aug 21 2014 11:54:54 GMT+0300 (GTB Daylight Time)
165	martar	312	2	Thu Aug 21 2014 11:54:54 GMT+0300 (GTB Daylight Time)

8.4 Γλωσσάρι

Πλαίσιο αλληλεπίδρασης χρήστη→Problem Context

Κινήσεις→Actions

Συνεδρία→Session

HMM→Hidden Markov Models

Κρυφά Μαρκοβιανά Μοντέλα→Hidden Markov Models

Σύνολο κινήσεων→Log

Αναζήτηση→Query

Κέρσορας→Mouse

Ανακάτεμα→Shuffle

Νέα αναζήτηση→New Query

Αδράνεια→Idle

Πνευματική εικόνα→Mental-Image

Εκπαίδευση μοντέλου→Model Training

Κεφάλαιο 9

Βιβλιογραφία

1) Ryen W.White, Resa A.Roth, *Exploratory Search – Beyond the Query – Response Paradigm,* Synthesis Lectures on information concepts, retrieval and services #3, 2009
2) Sergios Theodoridis, Aggelos Pikrakis, Konstantinos Koutroumbas, Dionisis Kavouras, *Introduction to pattern recognition a matlab approach ,*Athens, Paschalidis, 2006
3) Rob Capra, Gene Golovchinsky, Bill Kules, Ryen W.White, *Introduction to special issue on human–computer information retrieval,* Information Processing and Management : an International Journal , Volume 49 Issue 5, 2013
4) Carla Teixeira Lopes, *Context Features and their use in Information retrieval,* FDIA'09 Proceedings of the Third BCS-IRSG conference on Future Directions in Information Access pp36-42 *, 2009*
5) Kelly,D, Dumais,S. and Pedersen,J., *Evaluation challenges and directions for information seeking support systems,*IEEE computer society, Issue No.03 , 2009
6) White,R. Muresan,G. and Marchionini,G. , *Report on ACM SIGIR workshop on evaluating exploratory search systems,* SIGIR Forum, 40(2), pp52-60, 2006
7) Cronen-Townsend,S. Zhou,Y. and Croft,W.B, *Predicting query performance,* In proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retirieval, pp.299-306, 2002
8) Ingwersen,P. , *Information Retrieval Interaction,*London,UK:Taylor Graham,1992
9) Vakkari,P , *Task complexity, problem structure and information actions: Integrating studies on information seeking and retrieval,* Information Processing and Management,35(6)pp819-837, 1999
10) Taylor,R.S., *Question negotiation and information seeking inlibraries,* College and Research Libraries,29(3), pp178-194 , 1968
11) Schmidt,A., Beigl,M., Gellersen,HW., *There is more to context than location ,* Computers and Graphics 203-12 self, , 1998
12) White,R.W., Kules,B., Drucker,S., Schraefel,M.C., *Supporting exploratory search: Introduction to special section,* Communications of the ACM,49(4),pp.36-39, 2006

13) Pace,S., *A grounded theory of the flow experiences of web users,* International Journal of Human-Computer Studies,60(3),pp.327-363, 2004

14) Nagarajan Natarajan, Donghyuk Shin, Inderjit Dhillon, *Which App Will You Use Next? Collaborative Filtering with Interactional Context,* RecSys 13, Proceedings of the 7th ACM conference on Recommender Systems, 2013

15) Vincenzo Deufemia, Massimialiano Giordano, Giuseppe Polese,Luigi Marco Simonetti, *Exploiting Interaction Features in User Intent Understanding,* Web Technologies and applications , lecture notes in computer science, volume7808,pp506-517, 2013

16) Zhen Yue, Shuguang Han, Daqing He*, Modeling Search Processes using Hidden States in Collaborative Exploratory Web Search,* CSCW Proceedings of the 17th ACM conference on computer supported cooperative work and social computing, pp.820-830, 2014

17) Hao-Chuan Wang, *Modeling Idea Generation Sequences Using Hidden Markov Models* ,2014

18) Mark Stamp, *A revealing Introduction to Hidden Markov Models ,* CiteSeer, 2012

19) Andrew W. Moore, *Cross-Validation for Detecting and preventing overfitting ,* Carnegie Melon University,2001

20) Peter Hall, Jeff Racine, Qi Li, *Cross-Validation and the estimation of conditional probability densities ,*Journal of the Americal Statistical Assosiation, Vol99,no468, 2004

21) Tuukka Ruotsalo, Kumaripada et al, *Supporting Exploratory Search Tasks with Interactive User Modeling,* CIKM, Proceedings of the 22nd ACM international conference on information and knowledge management, 2013

22) Dey,A.K., G.D.Abowd and D.Salber,. *A conceptual framework and a toolkit for supporting the rapid prototyping and context-aware application*, Human-Computer Interaction, Volume 16, Issue 2 ,2001

23) Jeremy Pickens, Gene Golovchinsky, *Collaborative Exploratory Search,* SIGIR Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval ,2007

24) Gary Marchionini, *Exploratory Search: From finding to understanding,* Magazine communications of the ACM- Supporting exploratory search, Volume 49, Issue 4, 2006

25) Ralf Krestel, Gianluca Demartini, Eelco Herder, *Visual Interfaces for stimulating exploratory search* ,2011

26) Stefen Haun, Andreas Nurnberger, *Supporting Exploratory Search by User-Centered Interactive Data Mining,* JCDL Proceedings of the 11th annual international ACM/IEEE joint conference on Digital Libraries ,2006

27) Eric Fosler-Lussier, *Markov Models and Hidden Markov Models: A Brief Tutorial* , International computer science institute, 1998

28) Phil Blunsom, *Hidden Markov Models,* CiteSeer , 2004

29) Zoybin Ghahramani, *An introduction to Hidden Markov Models and Bayesian Networks* , AAAI, 2001

30) Jia Li, *Hidden Markov Model,* The Pennsylvania State University, 2011

31) Jonathan Chaffer, Karl Swedberg, *Learning jQuery,* Packt, 2013

32) Marijin Hoverbeke, *Eloquent JavaScript A modern introduction to programming* , London, 2011

33) Cody Lindley, *jQuery Enlightment* , New York, 2009

34) *http://www.mathworks.com/help/stats/hidden-markov-models-hmm.html*

35) *http://dev.mysql.com/doc/*

36) *http://www.w3schools.com/js/default.asp*

37) *http://www.w3schools.com/jquery/default.asp*

38) *http://learn.jquery.com/*

39) http://en.wikipedia.org/wiki/Spring_Framework,

40) https://maven.apache.org/