ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Μοντελοποίηση της πλοήγησης των χρηστών στον Παγκόσμιο Ιστό με χρήση μεθόδων Συμπερασμού Γραμματικών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Γεώργιος Λ. Κορφιάτης

Επιβλέπων : Τιμολέων Σελλής
Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2006

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Μοντελοποίηση της πλοήγησης των χρηστών στον Παγκόσμιο Ιστό με χρήση μεθόδων Συμπερασμού Γραμματικών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Γεώργιος Λ. Κορφιάτης

Επιβλέπων : Τιμολέων Σελλής
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 8η Μαρτίου 2006.

.....	.....	.....
Τ. Σελλής Καθηγητής Ε.Μ.Π.	Ι. Βασιλείου Καθηγητής Ε.Μ.Π.	Ν. Κοζύρης Επίκ. Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2006

.....

Γεώργιος Λ. Κορφιάτης

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

Περίληψη

Η έλλειψη δομής του Παγκόσμιου Ιστού και το πρόβλημα της υπερσυσσώρευσης πληροφοριών καθιστούν δυσχερή την πλοήγηση σε αυτόν. Στην παρούσα εργασία προτείνεται μία μέθοδος μοντελοποίησης της πλοήγησης των χρηστών στον Παγκόσμιο Ιστό, με στόχο τη διευκόλυνση του χρήστη μέσω της πρότασης ενδιαφερουσών σελίδων σε αυτόν. Για το σκοπό αυτό, επεκτάθηκαν οι μέθοδοι Συμπερασμού Γραμματικών Alergia και Blue Fringe με την εισαγωγή ενός επιπλέον κριτηρίου, που ελέγχει την ομοιότητα των ιστοσελίδων ως προς το περιεχόμενό τους. Επίσης χρησιμοποιήθηκε μία τεχνική μείωσης διαστασιμότητας πριν την εφαρμογή της επαγωγικής μεθόδου. Στο πλαίσιο του Συμπερασμού Γραμματικών, οι ιστοσελίδες θεωρούνται σύμβολα μίας πιθανοτικής κανονικής γραμματικής και οι αλληλουχίες σελίδων συμβολοσειρές της αντίστοιχης γλώσσας. Επιπλέον, το περιεχόμενο της κάθε σελίδας εκφράζεται με το διάνυσμα των λέξεων-κλειδιών της. Από τα δεδομένα χρήσης που λαμβάνονται από αρχεία καταγραφής μιας εταιρείας παροχής υπηρεσιών διαδικτύου κατασκευάζεται αρχικά μία δενδρική δομή, τέτοια ώστε κάθε σύνοδος χρήσης των υπαρχόντων δεδομένων να αντιστοιχεί σε ένα μονοπάτι στο δέντρο. Στη συνέχεια, η μέθοδος επάγει από το αρχικό δέντρο ένα γράφο μικρότερης τάξης, που επιχειρεί να μοντελοποιήσει την πλοήγηση των χρηστών. Αυτό επιτυγχάνεται με τη συγχώνευση καταστάσεων (κόμβων του γράφου) που είναι συμβατές τόσο ως προς τη χρήση (όμοιες μεταβάσεις) όσο και ως προς το περιεχόμενο (ομοιότητα του περιεχομένου των σελίδων). Ο τελικός γράφος χρησιμοποιείται για την πρόταση ενδιαφερόντων συνδέσμων σελίδων σε χρήστες που περιηγούνται στον Παγκόσμιο Ιστό.

Τα πειραματικά αποτελέσματα έδειξαν ότι η γνώση της σειράς με την οποία ένας χρήστης επισκέπτεται ορισμένες σελίδες του Παγκόσμιου Ιστού δε συμβάλλει στη διαδικασία πρότασης σελίδων, κάτι που οφείλεται στη μεγάλη ανομοιογένεια των δεδομένων χρήσης. Γενικά, φαίνεται ότι η πλοήγηση ενός χρήστη στον Παγκόσμιο Ιστό περιορίζεται κατά κύριο λόγο σε ένα σύνολο σελίδων της ίδιας θεματικής κατηγορίας, ενώ οι λίγες μεταβάσεις σε άλλες θεματικές κατηγορίες είναι δύσκολο να προβλεφθούν. Εκτιμάται πάντως ότι μία προσέγγιση που θα βασίζεται στην ομοιότητα περιεχομένου και θα χρησιμοποιεί τα δεδομένα χρήσης σε επιλεκτική βάση ενδέχεται να αποδίδει καλύτερα. Επίσης, προέκυψε ότι η μέθοδος που βασίζεται στην Blue Fringe αποδίδει καλύτερα, καθότι αυτή επιλέγει με πιο έξυπνο τρόπο τις καλύτερες συγχωνεύσεις καταστάσεων. Η μείωση διαστασιμότητας δε φάνηκε τέλος να βελτιώνει τη διαδικασία πρότασης σελίδων.

Λέξεις Κλειδιά: Μηχανική Μάθηση, Μοντελοποίηση της Χρήσης του Ιστού, Συμπερασμός Γραμματικών, Ανάκτηση Πληροφοριών, Ομαδοποίηση

Abstract

The lack of structure of the World Wide Web and the information overload problem make the navigation through it a difficult task. In this dissertation, a method that models the Web user navigation is presented, which aims at assisting a user by recommending pages. For that purpose, the Grammatical Inference methods Alergia and Blue Fringe have been extended, by introducing an extra criterion, which examines the content similarity of the Web pages. A dimensionality reduction technique has also been employed before applying the inductive method. In the context of Grammatical Inference, the Web pages are considered as symbols of a probabilistic regular grammar and the sequences of pages as strings of the respective language. Moreover, the content of each page is represented by a vector of its keywords. Based on the usage data that are taken from log files of an Internet Service Provider we construct initially a tree structure, such that each session of the existing usage data corresponds to a path on the tree. Then the method infers from the initial tree a graph of lower order, which attempts to model the user navigation. This is achieved by merging states (nodes of the graph) which are compatible with respect to both the usage (similar transitions) and the content (content similarity of the pages). The final graph is used for the recommendation of useful page links to users who navigate through the World Wide Web.

The experimental results showed that the knowledge of the order in which a user visits some pages on the Web does not contribute to the page recommendation process, due to the diversity of the usage data. It seems in general that the navigation of a user through the Web is mainly restricted to a set of pages of a single thematic category, while it is difficult for the few transitions to other thematic categories to be predicted. However, it is possible that an approach based on content similarity that would exploit the usage data selectively might perform better. Moreover, it turned out that the method based on Blue Fringe performs better, since it chooses the best merges in a more clever way. Finally, dimensionality reduction did not seem to improve the page recommendation process.

Keywords: Machine Learning, Web Usage Modeling, Grammatical Inference, Information Retrieval, Clustering

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα της διπλωματικής μου εργασίας, ερευνητή του Ε.Κ.Ε.Φ.Ε. «Δημόκριτος» Γιώργο Παλιούρα, για την καθοδήγησή του και την άριστη συνεργασία που είχαμε. Επίσης, ευχαριστώ τους Στασινό Κωνσταντόπουλο, Δημήτρη Πιερράκο και Νίκο Καραμπατζιάκη για τη βοήθεια που προσέφεραν. Ευχαριστώ τέλος τον καθηγητή Τίμο Σελλή που με παρότρυνε να αναλάβω διπλωματική εργασία στο Ε.Κ.Ε.Φ.Ε «Δημόκριτος» και ανέλαβε την εποπτεία της εργασίας αυτής.

Πίνακας περιεχομένων

1

Εισαγωγή

Τα τελευταία χρόνια η αλματώδης ανάπτυξη του Παγκόσμιου Ιστού (World Wide Web) έχει δημιουργήσει μεγάλες προσδοκίες, καθώς αποτελεί μια ανεκτίμητη πηγή πληροφοριών, χρήσιμη στην καθημερινή μας ζωή. Παράλληλα, ο αριθμός των χρηστών του Παγκόσμιου Ιστού έχει ανέλθει πλέον σχεδόν στο ένα δισεκατομμύριο. Η πλειοψηφία όμως αυτών των χρηστών δεν είναι ειδικοί του χώρου και δυσκολεύονται να αξιοποιήσουν τη σύγχρονη τεχνολογία. Το γεγονός αυτό σε συνδυασμό με την αυξανόμενη ποσότητα της πληροφορίας που συσσωρεύεται στον Ιστό καθιστά δύσκολο τον εντοπισμό της πραγματικά χρήσιμης πληροφορίας. Γι' αυτό, έχει γίνει πλέον επιτακτική η ανάγκη παροχής εξατομικευμένων υπηρεσιών στον Παγκόσμιο Ιστό (Web Personalization). Έχει καταστεί σαφές ότι μάλλον η ευκολότερη πρόσβαση σε μια υπηρεσία παρά η μεγάλη ποσότητα δεδομένων είναι αυτή που της προσδίδει προστιθέμενη αξία. Σε επίπεδο ιστοχώρου (website), η εξατομίκευση συνίσταται σε πολλές διαφορετικές λειτουργίες, μεταξύ των οποίων η προσαρμογή της ιστοσελίδας στα μέτρα του χρήστη καθώς και η καθοδήγησή του μέσα στον ιστοχώρο, προτείνοντάς του συνδέσμους (links) σε πιθανόν ενδιαφέρουσες σελίδες. Για την ανάπτυξη εξατομικευμένων υπηρεσιών ενός ιστοχώρου χρησιμοποιούνται συχνά τεχνικές από το πεδίο της Εξόρυξης Γνώσης από Δεδομένα (Data Mining). Όταν οι χρήστες του Παγκόσμιου Ιστού περιηγούνται σε έναν ιστοχώρο, η κίνησή τους αποθηκεύεται σε αρχεία καταγραφής (logs) του ιστοχώρου. Η γνώση που προκύπτει από την ανάλυση αυτών των δεδομένων χρήσης αξιοποιείται για την εξατομίκευση του ιστοχώρου.

Μία άλλη προσέγγιση του προβλήματος της υπερσυσσώρευσης δεδομένων αφορά στη διευκόλυνση του χρήστη στον εντοπισμό πληροφορίας από ολόκληρο τον Παγκόσμιο Ιστό. Καθώς ένας χρήστης περιηγείται στον Παγκόσμιο Ιστό, θα ήταν χρήσιμο να μπορούμε να του προτείνουμε συνδέσμους σε διάφορους ιστοχώρους που πιθανόν να τον ενδιαφέρουν. Για την ανάπτυξη ενός συστήματος που να προσφέρει αυτή τη λειτουργία χρειάζονται δεδομένα χρήσης ολόκληρου του Ιστού και όχι μόνο ενός συγκεκριμένου ιστοχώρου. Τέτοια δεδομένα υπάρχουν στα αρχεία καταγραφής των διακομιστών μεσολάβησης (proxy servers) των εταιρειών παροχής υπηρεσιών διαδικτύου (Internet Service Provider - ISP), τα οποία συλλέγονται συστηματικά κατά την περιήγηση των χρηστών-πελατών της εκάστοτε εταιρείας στο Διαδίκτυο. Ωστόσο, τα δεδομένα αυτά χαρακτηρίζονται από μεγάλη ανομοιογένεια, καθώς αντικατοπτρίζουν τη συμπεριφορά των χρηστών σε σχέση με ολόκληρο τον Παγκόσμιο Ιστό και όχι μόνο με έναν περιορισμένο ιστοχώρο. Η έλλειψη ομοιογένειας είναι γενικό χαρακτηριστικό του Παγκόσμιου Ιστού, καθώς αυτός αποτελείται από πάρα πολλές ιστοσελίδες που αναφέρονται σε ποικίλα, άσχετα μεταξύ τους θέματα. Το γεγονός αυτό καθιστά δυσχερή την εφαρμογή των συνήθων τεχνικών που βασίζονται μόνο στα δεδομένα χρήσης και που χρησιμοποιούνται για την εξατομίκευση ενός ιστοχώρου. Εκτιμάται πάντως ότι οι τεχνικές αυτές μπορούν να προσαρμοστούν κατάλληλα, εάν ληφθεί υπόψη και επιπλέον πληροφορία που να χαρακτηρίζει την ομοιότητα των διαφόρων ιστοσελίδων ως προς το περιεχόμενό τους. Η εισαγωγή της νέας αυτής παραμέτρου στοχεύει στο να αντισταθμίσει το έλλειμμα ομοιογένειας που παρατηρείται στα δεδομένα χρήσης του Ιστού με το να προσφέρει ένα επιπλέον μέτρο συσχέτισης των σελίδων.

1.1 Αντικείμενο της διπλωματικής

Στην παρούσα εργασία επιχειρείται να μοντελοποιηθεί η περιήγηση στον Παγκόσμιο Ιστό των χρηστών - πελατών μιας εταιρείας ISP. Στο πλαίσιο τεχνικών από το πεδίο της Εξόρυξης Γνώσης από Δεδομένα χρήσης του Παγκόσμιου Ιστού (Web Usage Mining) μελετάται η δυνατότητα συνδυασμού των δεδομένων χρήσης του Ιστού με πληροφορία για την ομοιότητα των ιστοσελίδων, με σκοπό την κατασκευή ενός γράφου που να περιγράφει τη συμπεριφορά των χρηστών ως προς την περιήγησή τους στον Παγκόσμιο Ιστό. Στον πυρήνα της διαδικασίας αυτής χρησιμοποιούνται τεχνικές από την περιοχή της Μηχανικής Μάθησης (Machine Learning), δηλαδή του επιστημονικού πεδίου που ασχολείται με την κατασκευή προγραμμάτων τα οποία βελτιώνονται αυτόματα με την εμπειρία που αποκτούν. Θεωρώντας κάθε αλληλουχία σελίδων που επισκέπτεται κάποιος χρήστης ως συμβολοσειρά μίας κανονικής γλώσσας, κάνουμε χρήση μεθόδων του Συμπερασμού Γραμματικών (Grammatical Inference), του κλάδου της Μηχανικής Μάθησης που ασχολείται με τον αυτόματο προσδιορισμό των κανόνων που διέπουν μια γλώσσα από προτάσεις που ανήκουν σε αυτή.

Αυτές οι μέθοδοι συνδυάζονται με τεχνικές από το πεδίο της Ανάκτησης Πληροφοριών (Information Retrieval), για να προκύψει το τελικό μοντέλο. Συγκεκριμένα, εφαρμόζεται μία διαδικασία που επάγει τον τελικό γράφο από μια αρχική δενδρική δομή. Για το σκοπό αυτό, συγχωνεύονται οι κόμβοι του γράφου που θεωρούνται συμβατοί τόσο ως προς τη χρήση (όμοιες μεταβάσεις) όσο και ως προς το περιεχόμενο (ομοιότητα περιεχομένου των ιστοσελίδων). Το εξαχθέν μοντέλο αποσκοπεί στο να προβλέψει την περαιτέρω πλοήγηση ενός χρήστη με βάση τις σελίδες που έχει μέχρι στιγμής επισκεφτεί ή εναλλακτικά να προτείνει στο χρήστη συνδέσμους σε πιθανόν ενδιαφέρουσες σελίδες.

1.2 Οργάνωση του τόμου

Το υπόλοιπο της εργασίας οργανώνεται ως εξής. Στο κεφάλαιο 2 παρατίθεται το απαραίτητο θεωρητικό υπόβαθρο. Συγκεκριμένα, αναπτύσσονται οι βασικές έννοιες από τις περιοχές της Εξόρυξης Γνώσης, της Μηχανικής Μάθησης, της Ομαδοποίησης καθώς και της Θεωρίας Τυπικών Γλωσσών. Το κεφάλαιο 3 ασχολείται με το πεδίο του Συμπερασμού Γραμματικών. Γίνεται μια γενική επισκόπηση και παρουσιάζονται αναλυτικά δύο αλγόριθμοι. Στο κεφάλαιο 4 παρουσιάζονται αρχικά οι εργασίες που έχουν σχετικό αντικείμενο και στη συνέχεια αναλύονται οι μέθοδοι μοντελοποίησης που αναπτύχθηκαν στο πλαίσιο της παρούσας εργασίας. Το κεφάλαιο 5 ασχολείται με την πειραματική αξιολόγηση των παραπάνω μεθόδων. Τέλος, στο κεφάλαιο 6 καταγράφονται τα συμπεράσματα που προέκυψαν από την εργασία αυτή και δίνονται κάποιες μελλοντικές κατευθύνσεις.

2

Υπόβαθρο

Στο κεφάλαιο αυτό παρατίθεται το απαραίτητο θεωρητικό υπόβαθρο. Η Ενότητα 1 ασχολείται με το πεδίο της Εξόρυξης Γνώσης. Στην Ενότητα 2 γίνεται μια εισαγωγή στη Μηχανική Μάθηση και ακολουθεί στην επόμενη ενότητα μια εισαγωγή στην Ομαδοποίηση. Τέλος, στην Ενότητα 4 δίνονται οι απαραίτητοι ορισμοί από τη Θεωρία Τυπικών Γλωσσών.

2.1 Εξόρυξη Γνώσης από Δεδομένα Χρήσης του Ιστού

2.1.1 Γενικά

Ο ολοένα και αυξανόμενος όγκος δεδομένων που συλλέγονται και αποθηκεύονται οδηγεί στην ανάπτυξη νέων, αποδοτικότερων μεθόδων επεξεργασίας και αξιοποίησης των δεδομένων αυτών. Οι κλασικές τεχνικές ανάκτησης αποθηκευμένων πληροφοριών δεν είναι πλέον επαρκείς. Συχνά, είναι χρήσιμη η αναζήτηση γνώσης που δεν είναι ρητά καταγεγραμμένη στα δεδομένα. Το ερευνητικό πεδίο που στοχεύει στην ανάπτυξη εργαλείων που αναζητούν κρυμμένα πρότυπα σε μεγάλες συλλογές δεδομένων λέγεται Εξόρυξη Γνώσης από Δεδομένα (Data Mining, Knowledge Discovery in Data) [PPP+03].

Καθώς ο Παγκόσμιος Ιστός μπορεί να θεωρηθεί ως μία αχανής συλλογή δεδομένων, έχει αναπτυχθεί ένας κλάδος της Εξόρυξης Γνώσης από Δεδομένα που χρησιμοποιεί τεχνικές από την περιοχή αυτή για την ανάλυση δεδομένων που έχουν συλλεχθεί από τον Παγκόσμιο Ιστό και για την εξαγωγή χρήσιμης γνώσης από αυτά. Ο κλάδος αυτός καλείται Εξόρυξη Γνώσης

από τον Παγκόσμιο Ιστό (Web Mining). Ένα μέρος της εργασίας σε αυτή την περιοχή εστιάζεται στην ανακάλυψη γνώσης σχετικά με τη συμπεριφορά των χρηστών του Ιστού, αναλύοντας δεδομένα χρήσης του Ιστού. Για τη διαδικασία αυτή χρησιμοποιείται ο όρος Εξόρυξη Γνώσης από Δεδομένα χρήσης του Παγκόσμιου Ιστού (Web Usage Mining). Αρχικά, ο στόχος αυτής της περιοχής ήταν να παρέχει υποστηρικτική γνώση σε συστήματα λήψης αποφάσεων που αφορούσαν θέματα διαχείρισης ενός ιστοχώρου ή ανάλυσης αγοράς. Καθώς όμως τα δεδομένα χρήσης απεικονίζουν την αλληλεπίδραση ανάμεσα στους χρήστες και τους ιστοχώρους, η Εξόρυξη Γνώσης από Δεδομένα χρήσης του Παγκόσμιου Ιστού συνιστά επίσης ένα χρήσιμο εργαλείο για την ανάπτυξη εξατομικευμένων υπηρεσιών στον Παγκόσμιο Ιστό. Η παρούσα εργασία εντάσσεται σε αυτό το πλαίσιο.

2.1.2 Στάδια της Διαδικασίας

Μία διαδικασία Εξόρυξης Γνώσης από Δεδομένα χρήσης του Παγκόσμιου Ιστού, όπως και κάθε διαδικασία Εξόρυξης Γνώσης από Δεδομένα, αποτελείται από τα εξής τέσσερα διαδοχικά στάδια: Συλλογή Δεδομένων, Προεπεξεργασία Δεδομένων, Ανακάλυψη Προτύπων και Εκμετάλλευση της Γνώσης. Παρακάτω παρουσιάζονται τα στάδια αυτά λαμβάνοντας υπόψη τη χρήση της διαδικασίας για εξατομίκευση στον Ιστό [PPP+03].

- **Συλλογή Δεδομένων:** Τα δεδομένα χρήσης που είναι απαραίτητα για τη διαδικασία μπορούν να συλλεχθούν από διάφορες πηγές: από έναν εξυπηρετητή Ιστού (web server), τοπικά από ένα χρήστη που προσπελαύνει κάποιον ιστοχώρο (πελάτης - client) ή από ενδιάμεσες πηγές, όπως από ένα διακομιστή μεσολάβησης (proxy server). Στην πρώτη περίπτωση, πρόκειται για δεδομένα χρήσης που έχουν αποθηκευτεί σε αρχεία καταγραφής (logs) του εξυπηρετητή. Σε αυτά καταγράφονται όλες οι σελίδες που ζητήθηκαν από τους χρήστες σε κάποιο χρονικό διάστημα καθώς και άλλες πληροφορίες, όπως η χρονική στιγμή κατά την οποία έγινε η αίτηση και η διεύθυνση διαδικτύου (IP address) του αιτούμενου. Στη δεύτερη περίπτωση, μπορεί ένα πρόγραμμα ενσωματωμένο σε μια ιστοσελίδα να εκτελεστεί στην πλευρά του πελάτη και να συλλέξει άμεσα πληροφορίες για τη συμπεριφορά του στον Ιστό. Στην τρίτη περίπτωση, πρόκειται συνήθως για αρχεία καταγραφής σε ένα διακομιστή μεσολάβησης μιας εταιρείας ISP. Αυτά μοιάζουν πολύ με τα αρχεία καταγραφής ενός εξυπηρετητή Ιστού, με τη διαφορά ότι καταγράφουν την κίνηση των πελατών της ISP σε ολόκληρο τον Ιστό και όχι σε ένα συγκεκριμένο ιστοχώρο. Για τις ανάγκες αυτής της εργασίας χρησιμοποιήθηκαν δεδομένα που ελήφθησαν από μία εταιρεία ISP.
- **Προεπεξεργασία Δεδομένων:** Σε αυτό το στάδιο επιχειρείται η προσαρμογή των ακατέργαστων δεδομένων που έχουν συλλεχθεί στο το προηγούμενο στάδιο σε μια ενιαία, συνεπή μορφή, έτσι ώστε να μπορούν να αξιοποιηθούν στο επόμενο στάδιο. Σε πρώτη

φάση πρέπει τα δεδομένα να καθαριστούν από πλεονάζουσα, άχρηστη πληροφορία. Όταν ένας χρήστης αιτείται μία ιστοσελίδα, γίνονται αυτόματα και επιπλέον αιτήσεις που αφορούν εικόνες ή βίντεο που πιθανόν υπάρχουν σε μια ιστοσελίδα. Αυτές οι αιτήσεις καταγράφονται στα αρχεία καταγραφής χωρίς να έχουν γίνει ρητά από το χρήση και συνεπώς θεωρούνται πλεονάζουσα πληροφορία. Έπειτα, πρέπει να γίνει η αναγνώριση των χρηστών, δηλαδή να διαπιστωθεί ποιες από τις καταγεγραμμένες αιτήσεις έγιναν από τον ίδιο χρήστη. Η πιο απλή προσέγγιση σε αυτό είναι να αντιστοιχιστεί σε κάθε διεύθυνση IP και ένας διαφορετικός χρήστης. Παρά την έλλειψη ακρίβειας αυτής της πρακτικής, αφού περισσότεροι χρήστες ενδέχεται να μοιράζονται την ίδια διεύθυνση IP, είναι τελικά αυτή που χρησιμοποιείται ευρύτερα. Τέλος, πρέπει να αναγνωριστούν οι σύνοδοι πλοήγησης (navigation sessions) στα δεδομένα χρήσης. Μία σύνοδος είναι η αλληλουχία των σελίδων που επισκέφθηκε ένας χρήστης κατά την περιήγησή του στον Ιστό. Για την αναγνώριση των συνόδων έχουν χρησιμοποιηθεί διάφορες ευριστικές τεχνικές. Μια συνηθισμένη πρακτική είναι η επιβολή ενός χρονικού ορίου. Αν κάποιος χρήστης παραμείνει αδρανής για κάποιο χρονικό διάστημα και μετά κάνει μία νέα αίτηση, τότε θεωρούμε ότι αυτή η νέα αίτηση εντάσσεται σε μία νέα σύνοδο. Αυτό το χρονικό όριο τίθεται συνήθως στα 30 λεπτά.

- **Ανακάλυψη Προτύπων:** Αυτό είναι το σημαντικότερο στάδιο της διαδικασίας, καθώς εδώ γίνεται η ανακάλυψη της επιθυμητής γνώσης από τα δεδομένα. Για το σκοπό αυτό, χρησιμοποιούνται τεχνικές από τη Μηχανική Μάθηση και τη Στατιστική. Για τις εφαρμογές εξατομίκευσης στον Ιστό, ως γνώση θεωρούνται κάποια πρότυπα που αντικατοπτρίζουν τη συμπεριφορά των χρηστών ως προς την περιήγησή τους στον Ιστό. Στη συνέχεια, αναφέρονται οι τέσσερεις βασικές προσεγγίσεις στην Ανακάλυψη Προτύπων:
  a) Ομαδοποίηση (Clustering): Με τις τεχνικές Ομαδοποίησης επιχειρείται η διαμέριση των δεδομένων σε ομάδες. Τα δεδομένα κάποιας ομάδας πρέπει να είναι σχετικά μεταξύ τους ως προς κάποιο μέτρο σύγκρισης αλλά αρκετά διαφορετικά με τα δεδομένα στις άλλες ομάδες. Για παράδειγμα, μπορούμε να ομαδοποιήσουμε ένα σύνολο ιστοσελίδων με βάση το περιεχόμενό τους.
  b) Ταξινόμηση (Classification): Ο στόχος μίας τεχνικής Ταξινόμησης είναι να αναγνωρίσει τα διακριτικά χαρακτηριστικά προκαθορισμένων κατηγοριών με βάση κάποια παραδείγματα, έτσι ώστε να μπορέσει στη συνέχεια να προβλέψει τη συμπεριφορά για νέες περιπτώσεις.
  c) Ανακάλυψη Συσχετίσεων (Association Discovery): Οι τεχνικές αυτές αποσκοπούν στην ανακάλυψη σχέσεων εξάρτησης μεταξύ δύο συνόλων αντικειμένων.

Χρησιμοποιούνται συνήθως για την ανακάλυψη συσχετίσεων ανάμεσα σε ιστοσελίδες που συναπαντώνται σε συνόδους πλοήγησης.

d) Ανακάλυψη Προτύπων Διαδοχής (Sequential Pattern Discovery): Εδώ εισάγεται το στοιχείο του χρόνου στη διαδικασία. Σκοπός είναι η αναγνώριση χρονικών προτύπων που παρατηρούνται συχνά στα δεδομένα. Η προσέγγιση αυτή είναι ιδιαίτερα χρήσιμη στην αναγνώριση προτύπων πλοήγησης σε δεδομένα χρήσης του Παγκόσμιου Ιστού. Για το σκοπό αυτό, χρησιμοποιούνται είτε ντετερμινιστικές τεχνικές, που καταγράφουν τη συμπεριφορά των χρηστών, είτε στοχαστικές, που αξιοποιούν την ακολουθία των επισκεφθέντων σελίδων για την πρόβλεψη επόμενων επισκέψεων.

Στην παρούσα εργασία γίνεται χρήση μίας στοχαστικής μεθόδου Ανακάλυψης Προτύπων Διαδοχής σε συνδυασμό με τεχνικές Ομαδοποίησης.

- **Εκμετάλλευση της Γνώσης:** Στο τελευταίο αυτό στάδιο γίνεται η ερμηνεία και η αξιολόγηση της γνώσης που έχει εξαχθεί και παρουσιάζεται σε κατανοητή μορφή. Επίσης, η γνώση αυτή αξιοποιείται σε συστήματα εξατομίκευσης στον Ιστό. Στην παρούσα εργασία το εξαχθέν μοντέλο χρησιμοποιείται για την εξατομίκευση της πλοήγησης στον Ιστό μέσω της πρότασης σελίδων σε χρήστες. Με βάση αυτό το στόχο αξιολογείται η επίδοση του μοντέλου.

2.2 Μηχανική Μάθηση

Η Μηχανική Μάθηση (Machine Learning) ασχολείται με την κατασκευή προγραμμάτων που βελτιώνονται αυτόματα με την εμπειρία που αποκτούν και έχει ποικίλες εφαρμογές στο πεδίο της Εξόρυξης Γνώσης από Δεδομένα, σε συστήματα διήθησης πληροφορίας και αλλού [Mit97]. Χρησιμοποιεί έννοιες από διάφορα επιστημονικά πεδία και κυρίως από τη στατιστική, την τεχνητή νοημοσύνη και τη θεωρία πληροφορίας. Ο όρος *μάθηση* χρησιμοποιείται εδώ με την έννοια ότι ένα πρόγραμμα μαθαίνει από την υπάρχουσα εμπειρία Ε σε σχέση με κάποια εργασία Τ που πρέπει να επιτελέσει και με κάποιο μέτρο απόδοσης Ρ, όταν η απόδοσή του στην εργασία Τ, όπως μετριέται από το Ρ, βελτιώνεται από την εμπειρία Ε. Για παράδειγμα, η εργασία ενός προγράμματος μπορεί να είναι η αναγνώριση χειρόγραφων λέξεων από εικόνες. Ως μέτρο αξιολόγησης μπορεί να θεωρηθεί σε αυτή την περίπτωση το ποσοστό των λέξεων που αναγνωρίστηκαν σωστά από ένα σύνολο δοκιμής. Τέλος, η εμπειρία θα είναι ένα σύνολο με χειρόγραφες λέξεις που είναι χαρακτηρισμένες εκ των προτέρων. Αυτό το σύνολο δεδομένων αξιοποιείται στην εκπαίδευση του προγράμματος.

Σε πολλές περιπτώσεις, το πρόγραμμα πρέπει να μάθει μία δυαδική συνάρτηση (για παράδειγμα, αν ο καιρός κάποια ημέρα είναι κατάλληλος για άθληση ή όχι ως προς ορισμένες μετεωρολογικές παραμέτρους). Τότε λέμε ότι αυτή η συνάρτηση αναπαριστά μία έννοια-

στόχο (target concept). Επίσης, τα αντικείμενα πάνω στα οποία ορίζεται η έννοια-στόχος λέγονται στιγμιότυπα (instances). Στο παράδειγμα, στιγμιότυπα είναι όλες οι ημέρες (όπως αναπαρίστανται ως διανύσματα των παραμέτρων που χαρακτηρίζουν τα στιγμιότυπα). Τα στιγμιότυπα που περιέχονται στην έννοια-στόχο καλούνται θετικά, σε αντίθετη περίπτωση αρνητικά. Έτσι, η εμπειρία που χρησιμοποιείται για τη μάθηση της έννοιας είναι ένα σύνολο S από παραδείγματα, δηλαδή ένα σύνολο από στιγμιότυπα μαζί με το χαρακτηρισμό αν ικανοποιούν την έννοια-στόχο. Το σύνολο αυτό λέγεται δείγμα εκπαίδευσης (training sample). Το υποσύνολο του S που περιέχει μόνο τα θετικά παραδείγματα συμβολίζεται ως S^+, ενώ αυτό που περιέχει μόνο τα αρνητικά S^-.

Ο σκοπός μιας διαδικασίας Μηχανικής Μάθησης είναι η προσέγγιση της έννοιας-στόχου από ένα σύνολο πιθανών λύσεων, που λέγονται υποθέσεις. Όμως, η μόνη διαθέσιμη πληροφορία που υπάρχει είναι το πεπερασμένο δείγμα εκπαίδευσης S. Συνεπώς, πρόκειται για μία διαδικασία επαγωγικού συλλογισμού (inductive reasoning), δηλαδή συμπερασμού μίας γενικότερης έννοιας από την ειδικότερη έννοια που αντιστοιχεί στο συγκεκριμένο δείγμα. Επομένως, το μόνο που είναι εγγυημένο είναι ότι η προσέγγιση της έννοιας-στόχου είναι ακριβής για τα δεδομένα εκπαίδευσης, ενώ για όλα τα άλλα στιγμιότυπα μπορούμε μόνο να το υποθέσουμε. Σε αυτό το ζήτημα αναφέρεται η θεμελιώδης παραδοχή της επαγωγικής μάθησης:

Αξίωμα: Οποιαδήποτε υπόθεση έχει βρεθεί να προσεγγίζει καλά την έννοια-στόχο σε ένα επαρκώς μεγάλο σύνολο παραδειγμάτων εκπαίδευσης, θα την προσεγγίζει καλά και σε άγνωστα στιγμιότυπα.

Παρόλα αυτά, τα δεδομένα εκπαίδευσης δεν είναι επαρκή. Έχει δειχθεί ότι, για να μπορέσει ένας αλγόριθμος μάθησης να κατατάξει άγνωστα στιγμιότυπα με κάποια λογική βάση, πρέπει να έχουν γίνει εκ των προτέρων κάποιες παραδοχές ως προς τη φύση της έννοιας-στόχου. Με αυτό τον τρόπο περιορίζεται ο χώρος των υποθέσεων στον οποίο αναζητείται η έννοια-στόχος και δίνεται μια λογική κατεύθυνση στη απαραίτητη γενίκευση πέρα από τα παραδείγματα εκπαίδευσης, έτσι ώστε να συναχθεί η έννοια-στόχος. Οι παραδοχές αυτές είναι γενικά διαφορετικές για κάθε αλγόριθμο μάθησης και δεν είναι απαραίτητο να γνωρίζουμε αν όντως ισχύουν. Το σύνολο των παραδοχών για κάποιο αλγόριθμο καλείται *επαγωγική προδιάθεση (inductive bias)* του αλγορίθμου. Πιο τυπικά, η επαγωγική προδιάθεση ενός αλγορίθμου L είναι οποιοδήποτε ελάχιστο σύνολο παραδοχών B, τέτοιο ώστε για κάθε έννοια-στόχο C και αντίστοιχα παραδείγματα εκπαίδευσης D_C, η δυαδική τιμή κατάταξης που ανατίθεται σε κάθε στιγμιότυπο x_i, $L(x_i, D_C)$, να συνεπάγεται λογικά από τα B, D_C και x_i:

$$\forall x_i \quad [(B \wedge D_C \wedge x_i) \vdash L(x_i, D_C)]$$

Με αυτό τον τρόπο επιτυγχάνεται η ισοδυναμία μιας διαδικασίας επαγωγικής μάθησης με μία παραγωγική (deductive) διαδικασία, δηλαδή με μια διαδικασία που με τυπικά ορθό τρόπο συμπεραίνει ειδικότερες έννοιες από γενικότερες.

2.3 Ομαδοποίηση

2.3.1 Γενικά

Η διαδικασία διαμέρισης ενός συνόλου φυσικών ή αφηρημένων αντικειμένων σε κλάσεις όμοιων αντικειμένων λέγεται Ομαδοποίηση (Clustering). Μία ομάδα (ή συστάδα - cluster) είναι μια συλλογή αντικειμένων τα οποία είναι όμοια μεταξύ τους και ανόμοια με τα αντικείμενα των άλλων ομάδων. Η ανομοιότητα των αντικειμένων προσδιορίζεται βάσει των τιμών των εκάστοτε χαρακτηριστικών που περιγράφουν τα αντικείμενα. Συνήθως το κάθε αντικείμενο εκφράζεται με ένα διάνυσμα των χαρακτηριστικών του και συνεπώς η ανομοιότητα μεταξύ δύο αντικειμένων μπορεί να προσδιοριστεί με ένα μέτρο απόστασης ανάμεσα στα αντίστοιχα διανύσματα. Το βασικό χαρακτηριστικό της ανάλυσης σε ομάδες που την αντιδιαστέλλει από την Ταξινόμηση (Classification) είναι ότι τα χαρακτηριστικά των ομάδων που σχηματίζονται δεν είναι γνωστά εκ των προτέρων. Συνεπώς, η Ομαδοποίηση αποτελεί χρήσιμο εργαλείο ανακάλυψης γνώσης, καθώς μπορεί να αποκαλύψει λανθάνουσες συσχετίσεις ανάμεσα σε περίπλοκα δεδομένα. Από τη στιγμή που έχει γίνει ομαδοποίηση των δεδομένων, μπορούν σε πολλές εφαρμογές τα δεδομένα μιας ομάδας να αντιμετωπιστούν συλλογικά κι όχι ατομικά. Έτσι επιτυγχάνεται μιας μορφής αφαίρεση (abstraction). Τεχνικές Ομαδοποίησης έχουν χρησιμοποιηθεί ευρέως σε διάφορες εφαρμογές, μεταξύ των οποίων η αναγνώριση προτύπων, η ανάλυση δεδομένων και η επεξεργασία εικόνας.

Τα βασικά σχεδιαστικά ζητήματα που ανακύπτουν σε μια διαδικασία Ομαδοποίησης είναι τα εξής [FB92]:

- Επιλογή των χαρακτηριστικών των αντικειμένων στη βάση των οποίων θα γίνει η ομαδοποίηση καθώς και της αναπαράστασής τους.
- Επιλογή του μέτρου συσχέτισης των αντικειμένων.
- Επιλογή της κατάλληλης μεθόδου ομαδοποίησης.

2.3.2 Μέτρα Συσχέτισης

Για να γίνει δυνατός ο διαμερισμός των αντικειμένων σε ομάδες, πρέπει να υπάρχει κάποιο μέτρο ποσοτικοποίησης του βαθμού συσχέτισης μεταξύ τους. Αυτό μπορεί να είναι ένα μέτρο απόστασης ή ένα μέτρο ομοιότητας ή ανομοιότητας. Ποιο μέτρο συσχέτισης θα επιλεχθεί εξαρτάται από το συγκεκριμένο πρόβλημα και είναι γενικά στη διακριτική ευχέρεια του

ερευνητή. Ένα κριτήριο για την επιλογή μετρικού είναι η μορφή των διαθέσιμων δεδομένων [HK01]. Τα χαρακτηριστικά των αντικειμένων αναπαρίστανται συνήθως ως ένα διάνυσμα μεταβλητών. Αυτές μπορεί να παίρνουν είτε συνεχείς τιμές σε κάποιο διάστημα είτε ένα πεπερασμένο αριθμό προκαθορισμένων τιμών που αντιστοιχούν σε κάποιες καταστάσεις. Αν υπάρχουν δύο μόνο προκαθορισμένες τιμές, η μεταβλητή λέγεται δυαδική.

Στην περίπτωση συνεχών μεταβλητών, υπάρχουν πολλά μετρικά απόστασης μεταξύ δύο αντικειμένων (μέτρα ανομοιότητας). Πολύ γνωστό είναι το μετρικό της Ευκλείδειας απόστασης. Έστω $x=\left(x_1,x_2,...,x_p\right)$ το διάνυσμα των χαρακτηριστικών για το πρώτο αντικείμενο και $y=\left(y_1,y_2,...,y_p\right)$ για το δεύτερο. Η Ευκλείδεια απόσταση υπολογίζεται ως

$$d\left(x,y\right)=\sqrt{\sum_i\left(x_i-y_i\right)^2}$$

Κάποια άλλα μετρικά υπολογίζουν την ομοιότητα των αντικειμένων αντί για την απόστασή τους. Αναφέρονται και ως μετρικά ομοιότητας προτύπων, διότι η τιμή τους αυξάνει όσο περισσότερο ταιριάζουν τα χαρακτηριστικά δύο αντικειμένων. Το πιο γνωστό είναι το μετρικό του συνημιτόνου:

$$COSINE\left(x,y\right)=\frac{\sum_i x_i y_i}{\sqrt{\left(\sum_i x_i^2\right)\left(\sum_i y_i^2\right)}}$$

Όταν οι μεταβλητές είναι δυαδικές, κατασκευάζουμε ένα πίνακα συνάφειας για κάθε ζεύγος αντικειμένων, στον οποίο καταγράφεται το πλήθος των χαρακτηριστικών που έχουν τιμή 1 και στα δύο αντικείμενα ή μόνο στο ένα ή σε κανένα.

		Αντικείμενο j	
		1	0
Αντικείμενο i	1	**q**	**r**
	0	**s**	**t**

Συνδυάζοντας τις τιμές του πίνακα, μπορούμε να ορίσουμε διάφορα μετρικά απόστασης. Για παράδειγμα, το παρακάτω μετρικό λαμβάνει υπόψη του το ποσοστό των ανόμοιων χαρακτηριστικών:

$$d\left(i,j\right)=\frac{r+s}{q+r+s+t}$$

2.3.3 Μέθοδοι Ομαδοποίησης

Υπάρχει ένας μεγάλος αριθμός αλγορίθμων ομαδοποίησης στη βιβλιογραφία. Η επιλογή του κατάλληλου εξαρτάται τόσο από τον τύπο των δεδομένων όσο και από τη συγκεκριμένη εφαρμογή. Οι μέθοδοι ομαδοποίησης χωρίζονται γενικά σε δύο μεγάλες κατηγορίες: τις μεθόδους διαμέρισης και τις ιεραρχικές [HK01].

- **Μέθοδοι Διαμέρισης:** Δοθείσης μίας βάσης από n αντικείμενα, μία μέθοδος διαμέρισης (partitioning method) χωρίζει τα δεδομένα σε k τμήματα, όπου $k \leq n$. Το κάθε τμήμα πρέπει να περιέχει τουλάχιστον ένα αντικείμενο και κάθε αντικείμενο πρέπει να ανήκει σε ακριβώς ένα τμήμα. Υπάρχουν ωστόσο και ορισμένες μέθοδοι που είναι λιγότερο αυστηροί σε αυτό. Σε κάθε περίπτωση, προς αποφυγή μιας εξαντλητικής εξέτασης όλων των πιθανών διαμερίσεων, οι μέθοδοι που χρησιμοποιούνται ευρέως είναι ευριστικές, κάνουν δηλαδή κάποιες εκ των προτέρων παραδοχές ως προς τη φύση του αποτελέσματος. Ο πιο γνωστός αλγόριθμος αυτής της κατηγορίας είναι ο αλγόριθμος των k μέσων τιμών (k-means). Σε αυτόν, κάθε ομάδα αναπαρίσταται από τη μέση τιμή των αντικειμένων της ομάδας, που μπορεί να θεωρηθεί ως το κέντρο βάρους της ομάδας. Με δεδομένο το k, η μέθοδος δημιουργεί μία αρχική διαμέριση. Στη συνέχεια, με μια επαναληπτική τεχνική μετακινεί τα αντικείμενα από ομάδα σε ομάδα με σκοπό τη βελτίωση της διαμέρισης.

Αλγόριθμος: k-means

Είσοδος: Το πλήθος των ομάδων k και η βάση με τα n αντικείμενα

Έξοδος: Σύνολο από k ομάδες που ελαχιστοποιεί το κριτήριο τετραγωνικού σφάλματος

Μέθοδος:

διάλεξε αυθαίρετα k αντικείμενα ως τα αρχικά κέντρα των ομάδων·

επανάλαβε

(επαν)ανάθεσε κάθε αντικείμενο στην ομάδα με την οποία το αντικείμενο είναι περισσότερο όμοιο ως προς την Ευκλείδεια απόσταση με βάση τη μέση τιμή των αντικειμένων στην ομάδα·

ενημέρωσε τις μέσες τιμές των ομάδων, δηλ. υπολόγισε τη μέση τιμή των αντικειμένων σε κάθε ομάδα (κέντρο βάρους)·

μέχρι να μη γίνονται πλέον αλλαγές.

Αλγόριθμος 2.1 Αλγόριθμος ομαδοποίησης k-means

- **Ιεραρχικές Μέθοδοι:** Μία μέθοδος ιεραρχικής ομαδοποίησης δημιουργεί μία ιεραρχική αποσύνθεση του δοθέντος συνόλου αντικειμένων. Έτσι, προκύπτει ένα δέντρο από ομάδες. Υπάρχουν δύο προσεγγίσεις στον τρόπο που γίνεται η αποσύνθεση. Κατά την πιο

συνηθισμένη, τη *συσσωρευτική (agglomerative)*, κάθε αντικείμενο τοποθετείται αρχικά σε μια δική του ομάδα και έπειτα συγχωνεύονται οι ομάδες σε μεγαλύτερες, μέχρις ότου ικανοποιηθεί μία συνθήκη τερματισμού. Πρόκειται δηλαδή για μια προσέγγιση από κάτω προς τα πάνω (bottom-up). Η δεύτερη προσέγγιση, η *διαιρετική (divisive)*, είναι μία προσέγγιση από πάνω προς τα κάτω (top-down) και λειτουργεί με αντίστροφο τρόπο από τη συσσωρευτική. Αρχικά, όλα τα αντικείμενα τοποθετούνται σε μία ομάδα και μετά αυτή υποδιαιρείται σε όλο και μικρότερες ομάδες, μέχρι να προκύψει ο επιθυμητός αριθμός ομάδων ή μέχρι η μέγιστη απόσταση μεταξύ δύο ομάδων να γίνει μεγαλύτερη από κάποιο κατώφλι. Το βασικό μειονέκτημα των ιεραρχικών μεθόδων είναι ότι, από τη στιγμή που θα έχει γίνει ένα βήμα συγχώνευσης ή διαίρεσης, αυτό δεν μπορεί πλέον να αναιρεθεί. Το γεγονός πάντως ότι δεν εξετάζεται ένα συνδυαστικό πλήθος επιλογών μειώνει σημαντικά το υπολογιστικό κόστος.

2.4 Τυπικές Γλώσσες

Στην ενότητα αυτή παρουσιάζονται οι βασικές έννοιες από τη θεωρία τυπικών γλωσσών που θα χρειαστούν στην παρούσα εργασία.

2.4.1 Γλώσσες και Γραμματικές

Ένα οποιοδήποτε μη κενό και πεπερασμένο σύνολο Σ αποτελούμενο από σύμβολα ονομάζεται *αλφάβητο*. Για παράδειγμα, το σύνολο {0, 1} είναι το δυαδικό αλφάβητο και το σύνολο {a, b, ..., y, z} το λατινικό αλφάβητο. Κάθε στοιχείο ενός αλφαβήτου Σ λέγεται *σύμβολο* του αλφαβήτου. Μια πεπερασμένη παράθεση από σύμβολα ονομάζεται *συμβολοσειρά*. Για παράδειγμα, τα a, abb, bcazaa είναι συμβολοσειρές του λατινικού αλφαβήτου. Οι συμβολοσειρές παριστάνονται συνήθως με μικρά ελληνικά γράμματα *α*, *β*, *γ*, κλπ. Η συμβολοσειρά που δεν περιέχει κανένα σύμβολο ονομάζεται *κενή* και παριστάνεται με *ε*. Ο αριθμός των συμβόλων που αποτελούν μια συμβολοσειρά *α* ονομάζεται *μήκος* της συμβολοσειράς και παριστάνεται με $|\alpha|$. Το σύνολο όλων των συμβολοσειρών που μπορούν να παραχθούν από ένα αλφάβητο Σ συμβολίζεται ως Σ*.

Έστω ένα αλφάβητο Σ. *Γλώσσα (language)* L επί του αλφαβήτου Σ ονομάζουμε ένα σύνολο συμβολοσειρών του Σ, δηλαδή ένα υποσύνολο του Σ*. Για παράδειγμα, μία γλώσσα επί του αλφαβήτου {a, b} μπορεί να είναι η $L_1 = \{ab^n \mid n \in \mathbb{N}\}$, που αποτελείται από όλες τις συμβολοσειρές που ξεκινούν με a και στη συνέχεια έχουν μηδέν ή περισσότερα b.

Μία *γραμματική (grammar)* G είναι ένα σύστημα παραγωγής συμβολοσειρών μίας γλώσσας. Εναλλακτικά, μπορούμε να θεωρήσουμε ότι γλώσσα της γραμματικής G είναι το σύνολο των

συμβολοσειρών L(G) που μπορούν να παραχθούν από τη γραμματική. Μία γραμματική ορίζεται από μία διατεταγμένη τετράδα της μορφής (T, N, P, S) όπου:

- T είναι ένα αλφάβητο, του οποίου τα μέλη ονομάζονται *τερματικά σύμβολα*. Παριστάνονται συνήθως με μικρά λατινικά γράμματα a, b, c, κλπ.
- N είναι ένα αλφάβητο, του οποίου τα μέλη ονομάζονται *μη τερματικά σύμβολα*. Τα T και N πρέπει να είναι ξένα μεταξύ τους. Παριστάνονται συνήθως με κεφαλαία λατινικά γράμματα A, B, C, κλπ.
- P είναι ένα πεπερασμένο σύνολο *κανόνων παραγωγής*. Οι κανόνες παραγωγής είναι διατεταγμένα ζεύγη (α, β) συμβολοσειρών του αλφαβήτου $T \cup N$ και συνήθως συμβολίζονται ως $\alpha \to \beta$. Το α λέγεται αριστερό μέλος του κανόνα και το β δεξιό μέλος.
- S είναι ένα στοιχείο του N, το οποίο ονομάζεται *αρχικό σύμβολο* της γραμματικής.

Η ιδέα είναι ότι το αλφάβητο T είναι το αλφάβητο της γλώσσας που θα παραχθεί· αντίθετα τα σύμβολα του αλφαβήτου N δε θα εμφανίζονται στην τελική γλώσσα, αλλά έχουν το ρόλο μεταβλητών στους κανόνες παραγωγής. Έτσι, με τη βοήθεια μιας γραμματικής μπορούμε να παραγάγουμε μια συμβολοσειρά ως εξής:

- Αρχίζουμε με τη συμβολοσειρά που περιέχει μόνο το S.
- Από την τρέχουσα συμβολοσειρά παράγουμε μια καινούργια αντικαθιστώντας κάποια υποσυμβολοσειρά της που αντιστοιχεί σε αριστερό μέλος κανόνα με το αντίστοιχο δεξιό μέλος. Επαναλαμβάνουμε όσες φορές χρειαστεί.
- Αν καταλήξουμε σε συμβολοσειρά που αποτελείται μόνο από τερματικά σύμβολα, τότε λέμε ότι αυτή παράγεται από τη γραμματική.

Για παράδειγμα, μία γραμματική G_1 που παράγει τη γλώσσα L_1 που ορίστηκε παραπάνω είναι η εξής: $T = \{\mathrm{a}, \mathrm{b}\}$, $N = \{S, B\}$, $P = \begin{Bmatrix} S \to \mathrm{aB} \\ \mathrm{B} \to \varepsilon \\ \mathrm{B} \to \mathrm{bB} \end{Bmatrix}$, S το αρχικό σύμβολο

Ανάλογα με την πολυπλοκότητα των κανόνων του συνόλου P, ο Noam Chomsky κατέταξε τις γραμματικές (και αντίστοιχα τις γλώσσες) σε μία ιεραρχία κλάσεων όπου κάθε κλάση είναι υποσύνολο της προηγούμενης:

- **Γραμματικές χωρίς περιορισμούς**: Στην κλάση αυτή ανήκουν όλες οι γραμματικές.
- **Γραμματικές με συμφραζόμενα (context-sensitive)**: Στην κλάση αυτή ανήκουν οι γραμματικές με κανόνες της μορφής $\alpha \to \beta$, όπου η συμβολοσειρά α περιέχει

τουλάχιστον ένα μη τερματικό σύμβολο και ισχύει $|\alpha| \le |\beta|$. Κατ' εξαίρεση επιτρέπεται ο κανόνας $S \to \varepsilon$.

- **Γραμματικές χωρίς συμφραζόμενα (context-free)**: Στην κλάση αυτή ανήκουν οι γραμματικές με κανόνες της μορφής $A \to \alpha$, όπου A μη τερματικό σύμβολο και α συμβολοσειρά.
- **Κανονικές (regular) γραμματικές**: Στην κλάση αυτή ανήκουν οι γραμματικές με κανόνες που έχουν μία από τις εξής μορφές: $A \to \mathrm{a}B$, $A \to \mathrm{a}$ ή $A \to \varepsilon$, όπου A και B μη τερματικά σύμβολα και a τερματικό.

Για παράδειγμα, η γραμματική G_1 είναι κανονική.

Στο πλαίσιο της εργασίας αυτής θα ασχοληθούμε με τις κανονικές γραμματικές που αποτελούν την απλούστερη και περισσότερο μελετημένη κλάση.

2.4.2 Αυτόματα

Μια αφηρημένη μηχανή *M* που παίρνει ως είσοδο συμβολοσειρές ενός ορισμένου αλφαβήτου Σ και δίνει ως έξοδο «ναι» ή «όχι», αν δηλαδή πρόκειται για έγκυρη συμβολοσειρά μιας δεδομένης γλώσσας L, λέγεται *αναγνωριστής* ή *αυτόματο (automaton)* της γλώσσας L. Έχει αποδειχτεί ότι κάθε κλάση γλωσσών μπορεί να αναγνωριστεί από ένα διαφορετικό τύπο αυτομάτου. Στη συνέχεια, θα επικεντρωθούμε στα πεπερασμένα αυτόματα που αναγνωρίζουν τις κανονικές γλώσσες.

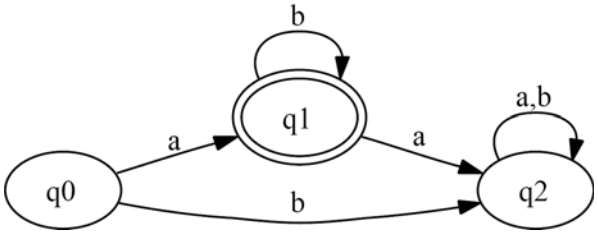
Ως είσοδος ενός πεπερασμένου αυτομάτου θεωρείται μία συμβολοσειρά. Κάθε σύμβολο που διαβάζεται οδηγεί σε αλλαγή κατάστασης του αυτομάτου. Στο τέλος της διαδικασίας μπορεί το αυτόματο να αποφανθεί αν η συμβολοσειρά ανήκει ή όχι στη γλώσσα, ανάλογα με την κατάσταση όπου έχει καταλήξει. Ανάλογα με τον τρόπο που πραγματοποιούνται οι μεταβάσεις από κατάσταση σε κατάσταση, τα πεπερασμένα αυτόματα διακρίνονται σε ντετερμινιστικά (deterministic finite automata - DFA) και μη ντετερμινιστικά (non-deterministic finite automata - NFA). Σε αντίθεση με τα δεύτερα, η λειτουργία των πρώτων καθορίζεται πλήρως από την είσοδό τους. Παρόλα αυτά, αποδεικνύεται ότι οι δύο αυτοί τύποι αυτομάτων είναι ισοδύναμοι.

Ένα ντετερμινιστικό πεπερασμένο αυτόματο ορίζεται τυπικά ως η πεντάδα $M = (Q, \Sigma, \delta, q_0, F)$ όπου:

- $Q = (q_0, q_1, ..., q_n)$ είναι ένα μη κενό πεπερασμένο σύνολο καταστάσεων.
- $\Sigma = (\mathrm{a}_1, \mathrm{a}_2, ..., \mathrm{a}_\mathrm{m})$ είναι το αλφάβητο του αυτομάτου
- $\delta : Q \times \Sigma \to Q$ είναι η συνάρτηση μετάβασης

- $q_0 \in Q$ είναι η αρχική κατάσταση
- $F \subseteq Q$ είναι το σύνολο των τελικών καταστάσεων

Για παράδειγμα, το DFA που αναγνωρίζει τη γλώσσα G_1 φαίνεται στο Σχήμα 2.1:



Σχήμα 2.1 Το DFA που αναγνωρίζει τη γλώσσα G_1

Σύνολο καταστάσεων $Q = \{q_0, q_1, q_2\}$, q_0 η αρχική κατάσταση, q_1 η τελική κατάσταση, αλφάβητο $\Sigma = \{\mathrm{a}, \mathrm{b}\}$ και συνάρτηση μετάβασης: $\begin{cases} \delta(q_0, \mathrm{a}) = q_1 & \delta(q_0, \mathrm{b}) = q_2 \\ \delta(q_1, \mathrm{a}) = q_2 & \delta(q_1, \mathrm{b}) = q_1 \\ \delta(q_2, \mathrm{a}) = q_2 & \delta(q_2, \mathrm{b}) = q_2 \end{cases}$.

2.4.3 Πιθανοτικές Γλώσσες

Οι *πιθανοτικές* ή *στοχαστικές γλώσσες (probabilistic, stochastic languages)* είναι χρήσιμες όταν οι συμβολοσειρές που παράγονται από μια γλώσσα δεν είναι το ίδιο πιθανό να εμφανιστούν και συνεπώς επιθυμούμε να έχουμε ένα μέτρο της συχνότητας εμφάνισης της κάθε συμβολοσειράς. Εδώ θα περιοριστούμε στις πιθανοτικές κανονικές γλώσσες, που είναι η επέκταση των κανονικών γλωσσών στο πεδίο των πιθανοτήτων.

Μία πιθανοτική κανονική γλώσσα μπορεί να παραχθεί από μία αντίστοιχη πιθανοτική κανονική γραμματική. Αυτή ορίζεται ως η διατεταγμένη δυάδα (G, p), όπου $G(T, N, P, S)$ μία κανονική γραμματική και $p : P \to [0,1]$ μία συνάρτηση πιθανότητας. Για κάποιον κανόνα $P_i \in P$, το $p(P_i)$ εκφράζει την πιθανότητα επιλογής του P_i ανάμεσα από όλους του κανόνες που έχουν το ίδιο αριστερό μέλος με τον P_i ως τον κανόνα που θα εφαρμοστεί.

Έτσι, η πιθανότητα εμφάνισης μιας συμβολοσειράς w ισούται με το γινόμενο των πιθανοτήτων των κανόνων που πρέπει να εφαρμοστούν για να παραχθεί.

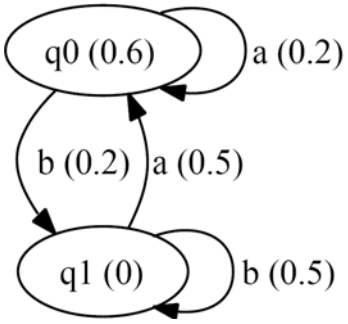
Το αυτόματο που αποφασίζει αν μια συμβολοσειρά w ανήκει σε μια πιθανοτική κανονική γλώσσα και υπολογίζει την πιθανότητά της λέγεται *στοχαστικό πεπερασμένο αυτόματο (stochastic finite automaton - SFA)*. Ένα SFA ορίζεται ως ένα πεπερασμένο αυτόματο $(Q, \Sigma, \delta, q_0, F)$ μαζί με τις συναρτήσεις $p : Q \times Q \times \Sigma \to [0,1]$ και $\pi_f : Q \to [0,1]$, όπου

$p(i,j,\mathrm{a})$ είναι η πιθανότητα μετάβασης από την κατάσταση q_i στην q_j όταν διαβαστεί στην είσοδο το σύμβολο a και $\pi_f(i)$ είναι η πιθανότητα η κατάσταση q_i να είναι τελική. Έτσι, για κάθε κατάσταση q_i πρέπει να ισχύει η παρακάτω συνθήκη:

$$\pi_f(i) + \sum_{q_j \in Q} \sum_{\mathrm{a} \in \Sigma} p(i,j,\mathrm{a}) = 1$$

Σε αντιστοιχία με το ντετερμινιστικό πεπερασμένο αυτόματο, ένα SFA είναι ντετερμινιστικό, όταν για κάθε κατάσταση q_i και σύμβολο εισόδου a υπάρχει μόνο μία κατάσταση q_j τέτοια ώστε $p(i,j,\mathrm{a}) \neq 0$. Όμως, ένα ντετερμινιστικό SFA δεν είναι ισοδύναμο με ένα μη ντετερμινιστικό SFA, σε αντίθεση με ό,τι ισχύει για την ισοδυναμία DFA και NFA.

Στο παρακάτω σχήμα φαίνεται ένα ντετερμινιστικό SFA που αναγνωρίζει συμβολοσειρές μια γλώσσας πάνω στο αλφάβητο $\{\mathrm{a}, \mathrm{b}\}$.



Σχήμα 2.2 Ένα SFA

Πάνω στις ακμές αναγράφεται σε παρένθεση η πιθανότητα της αντίστοιχης μετάβασης και μέσα στους κόμβους η πιθανότητα η κατάσταση να είναι τελική.

Στην εργασία αυτή γίνεται χρήση της έννοιας του στοχαστικού πεπερασμένου αυτομάτου.

3

Συμπερασμός Γραμματικών

Το κεφάλαιο αυτό ασχολείται με το αντικείμενο του Συμπερασμού Γραμματικών. Αρχικά παρουσιάζονται οι βασικές έννοιες. Στην Ενότητα 2 περιγράφονται ο χώρος αναζήτησης και αλγόριθμοι για το Συμπερασμό Κανονικών Γραμματικών και ακολουθεί μία ενότητα για το Συμπερασμό Γραμματικών χωρίς Συμφραζόμενα. Στις δύο τελευταίες ενότητες περιγράφονται αναλυτικά οι αλγόριθμοι Alergia και Blue Fringe.

3.1 Γενικά

Ένα από τα αντικείμενα της Μηχανικής Μάθησης είναι η αναγνώριση μιας άγνωστης συνάρτησης από ένα πεπερασμένο γνωστό σύνολο τιμών της. Μερικές φορές μπορεί να διατίθενται και αρνητικά παραδείγματα, δηλαδή τιμές που γνωρίζουμε ότι δεν παράγονται από τη συνάρτηση. Αυτή η διαδικασία αναγνώρισης λέγεται *Επαγωγικός Συμπερασμός (Inductive Inference)*. Αν θεωρήσουμε ότι το σύνολο των παραδειγμάτων αποτελείται από συμβολοσειρές ενός αλφαβήτου, τότε η έννοια που αναζητούμε είναι η γλώσσα από την οποία προέκυψαν τα δεδομένα μας. Συγκεκριμένα, υποθέτουμε ότι υπάρχει κάποια γραμματική G_0 που έχει παραγάγει τις συμβολοσειρές και στη συνέχεια επιδιώκουμε με βάση τις δεδομένες συμβολοσειρές να βρούμε μία γραμματική G που να είναι όσο κοντύτερα γίνεται στην G_0. Η διαδικασία αυτή λέγεται *Συμπερασμός Γραμματικών (Grammatical Inference)* ή *Επαγωγή Γραμματικών (Grammar Induction)* ([dlH05], [Dup97]). Για να οριστεί ακριβώς το πλαίσιο του προβλήματος, πρέπει να επιλεχθεί μία κατάλληλη

κλάση γραμματικών στην οποία υποθέτουμε ότι ανήκει η γραμματική-στόχος. Η περισσότερη έρευνα στο χώρο έχει γίνει πάνω στις κανονικές γραμματικές, καθώς αυτές είναι οι απλούστερες, διατηρώντας ωστόσο ικανοποιητική εκφραστικότητα. Εκτός αυτών, ερευνάται η επαγωγή και σε γραμματικές χωρίς συμφραζόμενα.

Ιδέες και τεχνικές από το πεδίο του Συμπερασμού Γραμματικών έχουν χρησιμοποιηθεί σε εφαρμογές διαφόρων πεδίων, όπως η Αναγνώριση Προτύπων, η Υπολογιστική Γλωσσολογία και η Βιοπληροφορική. Υπάρχει πάντως περιθώριο εύρεσης κατάλληλων προβλημάτων που να επιλύονται καλύτερα στο πλαίσιο του Συμπερασμού Γραμματικών παρά με άλλες τεχνικές της Μηχανικής Μάθησης [dlH05].

Ο Συμπερασμός Γραμματικών είναι ένα δύσκολο πρόβλημα και γίνεται δυσκολότερο όταν δεν υπάρχουν αρνητικά παραδείγματα, καθώς έχει αποδειχθεί [Gol67] ότι μια γραμματική δεν μπορεί να αναγνωριστεί σωστά μόνο από θετικά παραδείγματα. Συνεπώς, περιοριζόμαστε στην περίπτωση αυτή σε μία προσεγγιστική αναζήτηση.

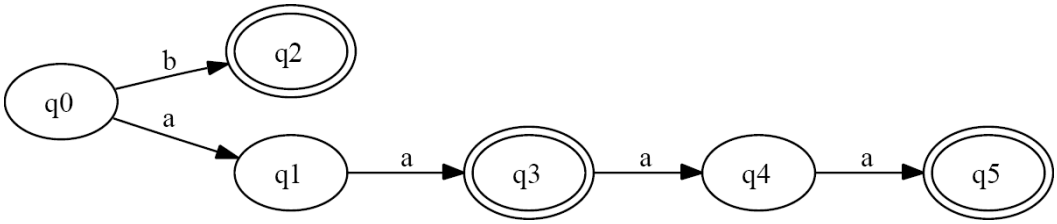
3.2 Συμπερασμός Κανονικών Γραμματικών

Είναι γνωστό ότι μία κανονική γραμματική μπορεί να αναπαρασταθεί με χρήση πεπερασμένων αυτομάτων. Ειδικότερα, υπάρχει θεωρητικά ένα DFA ανάμεσα σε αυτά που αναγνωρίζουν τη γλώσσα το οποίο είναι ελάχιστο. Αν και έχει δειχθεί ότι δεν υπάρχει αποδοτικός αλγόριθμος μάθησης που να μπορεί να αναγνωρίσει το ελάχιστο DFA που να είναι συνεπές με ένα σύνολο θετικών και αρνητικών παραδειγμάτων, είναι χρήσιμη η επαναδιατύπωση του προβλήματος επαγωγής πάνω σε κανονικές γραμματικές ως ένα πρόβλημα αναζήτησης στο χώρο των DFA.

3.2.1 Χώρος Αναζήτησης

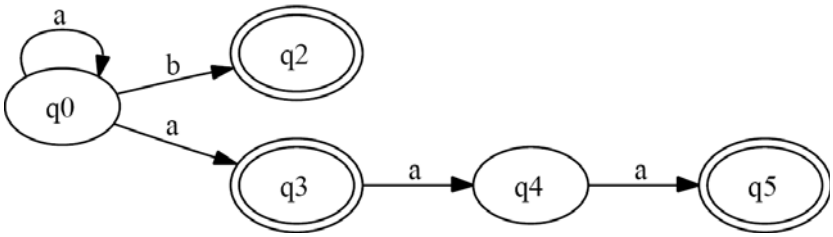
Ένα ζήτημα που πρέπει να ξεπεραστεί στο πρόβλημα της επαγωγής DFA είναι ότι ο χώρος αναζήτησης είναι άπειρος. Πολλές μέθοδοι συμπερασμού κανονικών γραμματικών κάνουν χρήση της έννοιας του *πλέγματος (lattice) από αυτόματα* για να περιορίσουν το χώρο αναζήτησης [PH00]. Αρχικά, το σύνολο των θετικών παραδειγμάτων S^+ χρησιμοποιείται για την κατασκευή ενός *δενδρικού αυτομάτου προθημάτων (prefix tree automaton - PTA)*. Αυτό είναι ένα DFA με ξεχωριστά μονοπάτια (modulo τα κοινά προθήματα) από την αρχική κατάσταση σε μία τελική κατάσταση για κάθε συμβολοσειρά του S^+. Έτσι, το PTA αναγνωρίζει μόνο τις συμβολοσειρές του S^+.

Για παράδειγμα, το παρακάτω PTA (Σχήμα 3.1) έχει κατασκευαστεί από το σύνολο $S^+ = \{\text{b, aa, aaaa}\}$.



Σχήμα 3.1 Ένα δενδρικό αυτόματο προθημάτων

Το πλέγμα ορίζεται τώρα ως το σύνολο όλων των δυνατών διαμερίσεων του συνόλου των καταστάσεων του PTA μαζί με μια σχέση που καθορίζει μια μερική διάταξη των στοιχείων του πλέγματος. Μία πιθανή διαμέριση είναι π.χ. η $\{\{q_0,q_1\},\{q_2\},\{q_3\},\{q_4\},\{q_5\}\}$. Κάθε στοιχείο του πλέγματος, δηλαδή κάθε διαμέριση των καταστάσεων του PTA, λέγεται *αυτόματο πηλίκου (quotient automaton)* και μπορεί να κατασκευαστεί από το PTA με συγχώνευση των καταστάσεων που ανήκουν στο ίδιο τμήμα της διαμέρισης. Το αυτόματο πηλίκου που προκύπτει για τη διαμέριση που αναφέρθηκε παραπάνω είναι το εξής:



Σχήμα 3.2 Ένα αυτόματο πηλίκου

Τα στοιχεία του πλέγματος διατάσσονται μερικώς από τη σχέση «καλύπτει». Λέμε ότι μια διαμέριση καλύπτει μια άλλη, αν η πρώτη παράγεται από τη συγχώνευση δύο ή περισσότερων καταστάσεων της δεύτερης. Για παράδειγμα, η διαμέριση $\{\{q_0,q_1,q_2\},\{q_3\},\{q_4,q_5\}\}$ καλύπτει τη $\{\{q_0,q_1\},\{q_2\},\{q_3\},\{q_4\},\{q_5\}\}$. Η ιδέα είναι ότι αν μια διαμέριση καλύπτει μια άλλη, τότε η γλώσσα την οποία αναπαριστά το αυτόματο πηλίκου της πρώτης είναι υπερσύνολο της γλώσσας που αναπαριστά το αυτόματο της δεύτερης. Δηλαδή το πρώτο αυτόματο είναι πιο γενικό από το δεύτερο. Από όλα τα στοιχεία του πλέγματος, το PTA είναι το πιο συγκεκριμένο, ενώ το παγκόσμιο DFA, που προκύπτει συγχωνεύοντας όλες τις καταστάσεις σε μία, είναι το πιο γενικό. Επομένως, ο χώρος αναζήτησης περιορίζεται στα στοιχεία του πλέγματος, δηλαδή στα αυτόματα που καλύπτουν το PTA και καλύπτονται από το παγκόσμιο DFA. Επειδή όμως το μέγεθός του παραμένει εκθετικό σε σχέση με τις καταστάσεις του PTA, τυπικές διαδικασίες αναζήτησης ξεκινούν από το PTA ή το παγκόσμιο DFA και κάνουν συγχωνεύσεις ή διασπάσεις καταστάσεων αντίστοιχα, για να δημιουργήσουν νέα στοιχεία του χώρου αναζήτησης.

3.2.2 Αλγόριθμοι Συμπερασμού

Υπάρχουν διάφοροι αλγόριθμοι που κάνουν αναζήτηση στο πλέγμα ([PH00], βιβλιογραφία στο [dlH05]). Μία στρατηγική είναι η αναζήτηση διπλής κατεύθυνσης από το PTA προς πιο γενικά αυτόματα και από το παγκόσμιο DFA προς τα πιο συγκεκριμένα. Όταν συναντηθούν τα δύο μέτωπα αναζήτησης, τότε θεωρείται ότι έχει εντοπιστεί το DFA-στόχος. Ο αλγόριθμος RPNI [OG92] ξεκινάει από το PTA και με κατά βάθος αναζήτηση στο πλέγμα επιλέγει τους κόμβους που θα συγχωνεύσει. Αν καταλήξει σε αυτόματο που αποδέχεται κάποιο αρνητικό αποτέλεσμα, τότε κάνει οπισθοχώρηση. Ωστόσο, οι τεχνικές αυτές δεν είναι αυξητικές, δηλαδή, για να μπορούν να αξιοποιηθούν νέα παραδείγματα, πρέπει η διαδικασία αναζήτησης να επαναληφθεί εξαρχής. Για την αντιμετώπιση αυτού του προβλήματος έχει προταθεί μία αυξητική έκδοση του RPNI [Dup96].

Για να αποφευχθεί η εξαντλητική αναζήτηση στο πλέγμα, που έχει συνδυαστικό μέγεθος, έχουν χρησιμοποιηθεί επίσης τεχνικές από την Τεχνητή Νοημοσύνη. Οι γενετικοί αλγόριθμοι [Dup94] ξεκινούν από το PTA και δημιουργούν ένα σύνολο από τυχαία επιλεχθέντα στοιχεία του πλέγματος. Στη συνέχεια, αξιοποιώντας κριτήρια βελτιστότητας τοποθετούν νέα στοιχεία στο σύνολο και τελικά επιλέγουν το καλύτερο από αυτά. Ένα άλλος αλγόριθμος, ο BIC [OS01], διενεργεί συγχωνεύσεις καταστάσεων και κάνει έξυπνη υπαναχώρηση αποφεύγοντας τον έλεγχο για αυτόματα των οποίων η ασυνέπεια μπορεί να συναχθεί εκ των προτέρων.

Υπάρχει ακόμη η δυνατότητα αξιοποίησης των νευρωνικών δικτύων στο συμπερασμό γραμματικών [GMC+92]. Συγκεκριμένα, χρησιμοποιούνται δίκτυα με ανατροφοδότηση και οι νευρώνες ανατροφοδότησης αντιστοιχούν στην τρέχουσα κατάσταση του αυτομάτου. Η τεχνική αυτή δίνει τη δυνατότητα μάθησης από λίγα μόνο παραδείγματα, ενώ ταυτόχρονα μπορεί εύκολα να κλιμακωθεί για μεγαλύτερα προβλήματα.

Όλες οι παραπάνω μέθοδοι κάνουν χρήση τόσο θετικών όσο και αρνητικών παραδειγμάτων. Ωστόσο, συχνά σε πρακτικές εφαρμογές υπάρχουν μόνο θετικά παραδείγματα. Σε αυτή την περίπτωση, η εκμάθηση της γραμματικής είναι εν γένει πιο δύσκολη. Ένας αλγόριθμος που συμπεραίνει μια γραμματική μόνο από θετικά παραδείγματα είναι ο k-TSSI [GV90], ο οποίος δεν κάνει χρήση του πλέγματος, αλλά κατασκευάζει ένα αυτόματο τέτοιο ώστε όλες οι συμβολοσειρές που έχουν κοινά τα τελευταία k-1 σύμβολα (για κάποιο επιλεγμένο k) να δείχνουν στην ίδια κατάσταση. Για την αντιμετώπιση της έλλειψης αρνητικών παραδειγμάτων χρησιμοποιούνται συχνά και ευριστικές τεχνικές. Για παράδειγμα, ο ευριστικός αλγόριθμος ECGI [RV88] κατασκευάζει ένα αυτόματο αξιοποιώντας την ιδέα της διόρθωσης σφαλμάτων.

Οι παραπάνω προσεγγίσεις στο πρόβλημα της μάθησης μόνο από θετικά παραδείγματα συχνά δεν έχουν ικανοποιητικά αποτελέσματα. Ωστόσο, με την εισαγωγή πιθανοτήτων στο μοντέλο

η διαδικασία μάθησης μπορεί να βελτιωθεί σημαντικά. Μία τέτοια προσέγγιση κάνει χρήση μαρκοβιανών μοντέλων. Μία άλλη προσέγγιση χρησιμοποιεί πιθανοτικές κανονικές γραμματικές ή αντίστοιχα στοχαστικά πεπερασμένα αυτόματα (SFA). Ο αλγόριθμος Alergia [CO94] ξεκινά από το PTA και κάνει συγχωνεύσεις, όπως και ο RPNI. Όμως, εδώ η σύγκριση των κόμβων γίνεται με στατιστικό τρόπο λαμβάνοντας υπόψη τις πιθανότητες μετάβασης. Έχουν αναπτυχθεί επίσης διάφορες παραλλαγές του Alergia, όπως ο αλγόριθμος MDI [TDH00], που αποφασίζει με διαφορετικό τρόπο τη συγχώνευση των καταστάσεων. Ο αλγόριθμος Blue Fringe [LPP98] βασίζεται στον Alergia, αλλά χρησιμοποιεί την ιδέα ότι πρέπει να γίνονται πρώτα οι συγχωνεύσεις για τις οποίες υπάρχουν περισσότερες θετικές ενδείξεις. Οι αλγόριθμοι Alergia και Blue Fringe θα περιγραφούν αναλυτικά στη συνέχεια.

3.3 Συμπερασμός Γραμματικών χωρίς Συμφραζόμενα

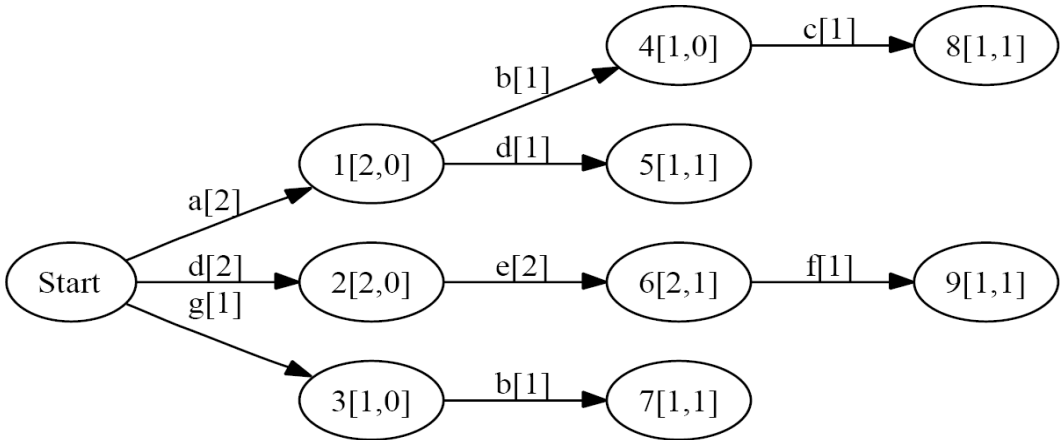
Η εκμάθηση ολόκληρης της κλάσης των γλωσσών χωρίς συμφραζόμενα φαίνεται να είναι υπολογιστικά αδύνατη ανεξαρτήτως μοντέλου μάθησης [dlH05]. Ωστόσο, για ορισμένες ειδικές περιπτώσεις γλωσσών μπορούν να επεκταθούν τα αποτελέσματα από την κλάση των κανονικών γλωσσών καθιστώντας έτσι εφικτό το συμπερασμό τους. Στη βιβλιογραφία υπάρχει π.χ. ένα πλήθος αποτελεσμάτων αναφορικά με τις λεγόμενες άρτιες γραμμικές γλώσσες [Tak88], ενώ οι γραμμικές γλώσσες έχουν ερευνηθεί και για την περίπτωση ύπαρξης μόνο θετικών παραδειγμάτων [KMT97]. Επίσης, έχουν προταθεί τεχνικές που χρησιμοποιούν γενετικούς αλγορίθμους [SK99]. Μία άλλη προσέγγιση του ζητήματος κάνει χρήση των δενδρικών αυτομάτων [FB75], που είναι μία επέκταση των DFA για δέντρα αντί για συμβολοσειρές. Τα δενδρικά αυτόματα παρέχουν έναν τρόπο σύνδεσης των αυτομάτων με τις γραμματικές χωρίς συμφραζόμενα. Με αυτό τον τρόπο μπορεί να επιτευχθεί ο συμπερασμός τους με χρήση μεθόδων επαγωγής κανονικών γλωσσών. Τέλος, παρουσιάζει αρκετό ενδιαφέρον ο συμπερασμός πιθανοτικών γραμματικών χωρίς συμφραζόμενα, οι οποίες εμφανίζονται σε πολλές πρακτικές εφαρμογές, π.χ. [WA02]. Το πρόβλημα παραμένει ανοιχτό, καθώς φαίνεται να είναι αρκετά πιο δύσκολο από την επαγωγή ενός SFA.

3.4 Ο Αλγόριθμος Alergia

Ο αλγόριθμος Alergia [CO94] αποτελεί τη βασική μέθοδο εκμάθησης ενός στοχαστικού αυτομάτου. Αρχικά, ο αλγόριθμος κατασκευάζει από το σύνολο των θετικών παραδειγμάτων S^+ ένα πιθανοτικό δενδρικό αυτόματο προθημάτων (probabilistic prefix tree automaton - PPTA), που είναι η στοχαστική επέκταση του PTA. Σε ένα PPTA, σε κάθε κατάσταση q συμπεριλαμβάνεται ο αριθμός των συμβολοσειρών $C(q)$ που φτάνουν σε αυτή καθώς και ο

αριθμός των συμβολοσειρών $C(q,\#)$ που τερματίζουν σε αυτή. Επίσης, κάθε μετάβαση με σύμβολο a που ξεκινά από την κατάσταση q χαρακτηρίζεται από το πλήθος των συμβολοσειρών $C(q,\mathrm{a})$ που τη χρησιμοποιούν. Έτσι, η πιθανότητα η κατάσταση q να είναι τερματική είναι $C(q,\#)/C(q)$ και η πιθανότητα χρήσης της μετάβασης για το σύμβολο a είναι $C(q,\mathrm{a})/C(q)$.

Παρακάτω (Σχήμα 3.3) φαίνεται το PPTA που προκύπτει από το δείγμα $S^+ = \{\mathrm{abc, de, ad, def, gb}\}$. Μέσα σε κάθε κόμβο σημειώνεται το όνομα της κατάστασης (1, 2, ...) και στην αγκύλη οι τιμές $C(q)$ και $C(q,\#)$. Σε κάθε μετάβαση σημειώνεται το αντίστοιχο σύμβολο που διαβάζεται στην είσοδο και σε αγκύλη η τιμή $C(q,\mathrm{a})$. Οι κόμβοι του δέντρου αριθμούνται κατά πλάτος σύμφωνα με τη λεξικογραφική σειρά των προθημάτων τους.



Σχήμα 3.3 Το PPTA που αντιστοιχεί στο δείγμα

Μετά την κατασκευή του PPTA, ο αλγόριθμος συγκρίνει με λεξικογραφική σειρά κάθε κόμβο με όλους τους προηγούμενούς του και τον συγχωνεύει με τον πρώτο συμβατό κόμβο που θα βρει. Σχηματικά παρουσιάζεται ο Alergia στον Αλγόριθμο 3.1.

Για τον προσδιορισμό της συμβατότητας των καταστάσεων αξιοποιούνται οι καταχωρημένες πιθανότητες μετάβασης και τερματισμού. Η ιδέα είναι ότι, αν το δείγμα περιείχε όλες τις συμβολοσειρές της γλώσσας, τότε δύο καταστάσεις που θα αντιστοιχούσαν στην ίδια κατάσταση του αυτομάτου-στόχου θα είχαν για κάθε σύμβολο τις ίδιες μεταβάσεις, τόσο ως προς την πιθανότητα μετάβασης όσο και ως προς την κατάσταση-προορισμό. Επίσης, θα είχαν την ίδια πιθανότητα να είναι τερματικές καταστάσεις. Σχηματικά:

$$q_i \equiv q_j \rightarrow \left(C(q_i,\#) = C(q_j,\#) \wedge \forall \mathrm{a} \in \Sigma : \left(\frac{C(q_i,\mathrm{a})}{\mathrm{C(q_i)}} = \frac{C(q_j,\mathrm{a})}{\mathrm{C(q_j)}} \wedge \delta(q_i,\mathrm{a}) \equiv \delta(q_j,\mathrm{a}) \right) \right)$$

Αλγόριθμος: Alergia

Είσοδος: S^+ : Σύνολο θετικών παραδειγμάτων

α : 1 - βαθμός εμπιστοσύνης

Έξοδος: Ένα στοχαστικό DFA

```
begin
    A = PPTA from S+
    for j = successor(firstnode(A)) to lastnode(A)
         for i = firstnode(A) to j
              if compatible(qi , qj , α)
                   merge(A, qi , qj)
                   break inner loop
    return A
end
```

Αλγόριθμος 3.1 Η μέθοδος Alergia

Επειδή όμως τα δεδομένα που διαθέτουμε είναι περιορισμένα και επιδέχονται στατιστικές διακυμάνσεις, οι παραπάνω ισότητες πρέπει να προσεγγιστούν μέσα σε ένα διάστημα εμπιστοσύνης.

Όταν κάνουμε n πειράματα Bernoulli (ρίχνουμε n φορές ένα νόμισμα) με γνωστή πιθανότητα h για το ένα ενδεχόμενο (π.χ. κεφαλή) και παρατηρήσουμε f φορές το ενδεχόμενο αυτό (φέρουμε f φορές κεφαλή), τότε το διάστημα εμπιστοσύνης για την τυχαία μεταβλητή δίνεται από το λεγόμενο φράγμα Hoeffding:

$$\left|h - \frac{f}{n}\right| < \sqrt{\frac{1}{2n}\ln\frac{2}{\alpha}} \quad \text{με πιθανότητα } p > 1-\alpha$$

Επεκτείνοντας την παραπάνω ιδέα για δύο ανεξάρτητες τυχαίες μεταβλητές Bernoulli, ο αλγόριθμος θεωρεί ότι δύο καταστάσεις είναι διαφορετικές ως προς κάποια μετάβαση ή ως προς το αν είναι τερματικές, όταν ικανοποιείται η ανισότητα του Αλγορίθμου 3.2.

Χρησιμοποιώντας τον ορισμό αυτό, η συμβατότητα δύο καταστάσεων υπολογίζεται όπως στον Αλγόριθμο 3.3. Όπως φαίνεται σε αυτό, ο αλγόριθμος θεωρεί δύο καταστάσεις συμβατές, μόνο όταν δε διαφέρουν για καμία μετάβαση ούτε ως προς την πιθανότητα να είναι τερματικές. Επίσης, όταν δύο καταστάσεις δεν είναι διαφορετικές ως προς κάποια μετάβαση, γίνεται επιπλέον έλεγχος συμβατότητας για τις δύο καταστάσεις όπου καταλήγει αντίστοιχα η μετάβαση. Αυτή η αναδρομή δεν κινδυνεύει πάντως να γίνει ατέρμονη, καθώς η σειρά που γίνονται οι συγχωνεύσεις εξασφαλίζει το ότι τουλάχιστον η μία ελεγχόμενη κατάσταση είναι πάντα ρίζα ενός υποδέντρου.

Αλγόριθμος: different(n, n', f, f', α)

Είσοδος: n, n' : πλήθος συμβολοσειρών που φτάνουν στον αντίστοιχο κόμβο

f, f' : πλήθος συμβολοσειρών που τερματίζουν στον αντίστοιχο κόμβο ή ακολουθούν μία δεδομένη μετάβαση

α : 1 - βαθμός εμπιστοσύνης

Έξοδος: Λογική τιμή

begin

return $\left|\frac{f}{n}-\frac{f'}{n'}\right| > \sqrt{\frac{1}{2}\log\frac{2}{\alpha}}\left(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{n'}}\right)$

end

Αλγόριθμος 3.2 Υπολογισμός ομοιότητας κόμβων στον Alergia

Αλγόριθμος: compatible(q_i, q_j, α)

Είσοδος: q_i, q_j : καταστάσεις

α : 1 - βαθμός εμπιστοσύνης

Έξοδος: Αληθές, αν οι δύο καταστάσεις είναι συμβατές

```
begin
    if different(C(q_i), C(q_j), C(q_i,#), C(q_j,#), α)
        return false
    foreach a∈Σ
        if different(C(q_i), C(q_j), C(q_i,a), C(q_j,a), α)
            return false
        if not compatible(δ(q_i,a), δ(q_j,a), α)
            return false
    return true
end
```

Αλγόριθμος 3.3 Υπολογισμός συμβατότητας κόμβων στον Alergia

Εάν δύο καταστάσεις βρεθούν συμβατές, τότε συγχωνεύονται όπως φαίνεται στον Αλγόριθμο 3.4. Ο νέος κόμβος διαθέτει την ένωση των μεταβάσεων των κόμβων από τον οποίο προήλθε, με πιθανότητες μετάβασης ίσες με το άθροισμα των αντίστοιχων πιθανοτήτων. Αξίζει να σημειωθεί ότι αναδρομικά συγχωνεύονται και οι καταστάσεις όπου καταλήγουν οι μεταβάσεις από τις αρχικές καταστάσεις. Αυτό γίνεται για να διατηρηθεί ο ντετερμινισμός του γράφου. Η συμβατότητα των καταστάσεων αυτών είναι εγγυημένη λόγω του

αναδρομικού ελέγχου που προηγήθηκε. Σημειώνεται ότι ο νέος κόμβος παίρνει στην αρίθμηση τη θέση του κόμβου με το μικρότερο αύξοντα αριθμό.

Αλγόριθμος: merge(A, q_i , q_j)

Είσοδος: A : το PPTA

q_i , q_j : καταστάσεις προς συγχώνευση

begin

A = A − $\{q_i, q_j\}$

A = A ∪ q′

ID of q′ = min(IDs of q_i, q_j)

$C(q') = C(q_i) + C(q_j)$

$C(q', \#) = C(q_i, \#) + C(q_j, \#)$

for each a ∈ Σ

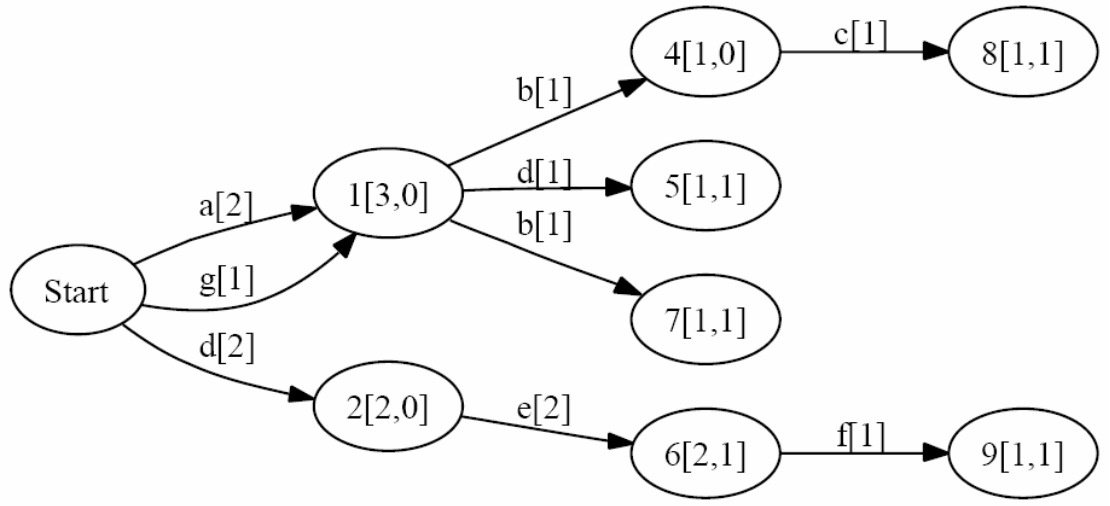
$C(q', a) = C(q_i, a) + C(q_j, a)$

merge(A, $\delta(q_i, a), \delta(q_j, a)$)

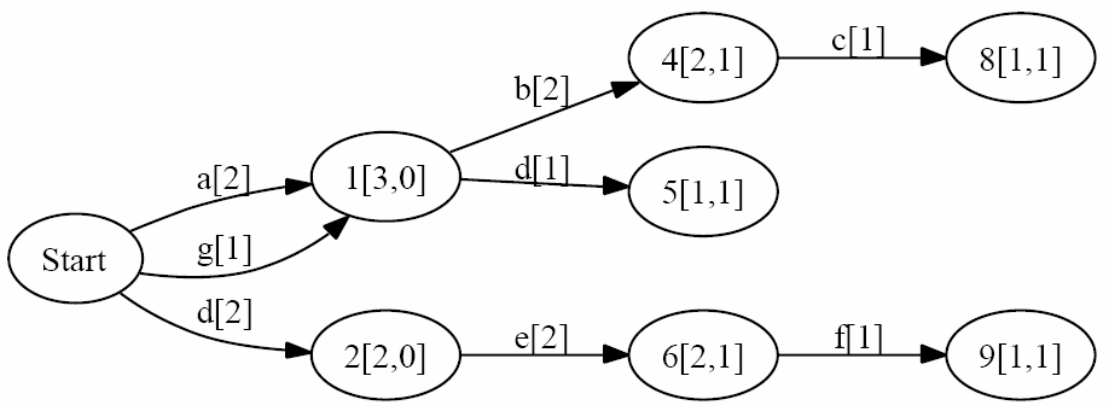
end

Αλγόριθμος 3.4 Διαδικασία συγχώνευσης κόμβων στον Alergia

Για παράδειγμα, αν βρεθούν συμβατοί οι κόμβοι 1 και 3 του Σχήματος 3.3, τότε οι δύο κόμβοι συγχωνεύονται (Σχήμα 3.4). Επειδή όμως ο κόμβος 1 έχει τώρα δύο διαφορετικές μεταβάσεις για το σύμβολο b, ακολουθεί η συγχώνευση των κόμβων 4 και 7, για να διατηρηθεί η ντετερμινιστικότητα του γράφου (Σχήμα 3.5).



Σχήμα 3.4 Το αυτόματο μετά τη συγχώνευση των 1 και 3



Σχήμα 3.5 Το αυτόματο μετά τη συγχώνευση των 4 και 7

Ο αλγόριθμος Alergia εγγυάται ότι θα βρει το SFA-στόχο στο όριο και η πολυπλοκότητά του είναι στη χειρότερη περίπτωση $O\left(\left\|S^+\right\|^3\right)$. Στην πράξη όμως ο χρόνος που απαιτείται είναι γραμμικός ως προς το $\left\|S^+\right\|$ και γενικά μπορεί να λειτουργήσει ακόμη και με σχετικά μικρό αριθμό παραδειγμάτων.

3.5 Ο Αλγόριθμος Blue Fringe

Το 1997 αναπτύχθηκε στο πλαίσιο του διαγωνισμού εκμάθησης DFA "Abbadingo One" η ιδέα της συγχώνευσης καταστάσεων οδηγούμενης από ενδείξεις (evidence-driven state merging - EDSM) [LPP98]. Σύμφωνα με αυτή την ευριστική προσέγγιση πρέπει να γίνονται πρώτα οι συγχωνεύσεις για τις οποίες υπάρχουν περισσότερες ενδείξεις ότι είναι σωστές. Για το σκοπό αυτό απαιτείται ένα μετρικό βαθμολόγησης των συγχωνεύσεων. Ωστόσο ο εξαντλητικός έλεγχος όλων των πιθανών συγχωνεύσεων εισάγει μεγάλη πολυπλοκότητα στη διαδικασία μάθησης. Γι' αυτό η στρατηγική Blue Fringe χρησιμοποιεί μια ευριστική προσέγγιση για τη μείωση του πλήθους των ελέγχων. Η ιδέα είναι ότι η μέθοδος διατηρεί ένα σύνολο κόμβων που έχουν ήδη ελεγχθεί και πλέον δεν μπορούν να συγχωνευτούν μεταξύ τους (κόκκινοι κόμβοι) και ένα σύνολο κόμβων που είναι υποψήφιοι προς συγχώνευση με κάποιον κόκκινο (μπλε κόμβοι).

Η διαδικασία ξεκινά με το PTA και χρωματίζεται η ρίζα κόκκινη. Τα παιδιά της χρωματίζονται μπλε και οι υπόλοιποι κόμβοι του δέντρου άσπροι. Σε όλη τη διαδικασία πρέπει να διατηρούνται οι παρακάτω αναλλοίωτες:

- Υπάρχει ένας αυθαίρετα συνδεδεμένος γράφος από αμοιβαία μη συγχωνεύσιμους κόκκινους κόμβους.
- Όλα τα παιδιά ενός κόκκινου κόμβου είναι ή κόκκινα ή μπλε.
- Οι μπλε κόμβοι είναι ρίζες απομονωμένων δέντρων.

Επίσης οι δυνατές πράξεις περιορίζονται στις παρακάτω:

- Υπολογισμός της βαθμολογίας συγχώνευσης ενός ζεύγους κόκκινου / μπλε κόμβου.
- Προαγωγή ενός μπλε κόμβου σε κόκκινο, αν δεν είναι συμβατός με κανένα κόκκινο.
- Συγχώνευση ενός μπλε κόμβου με ένα κόκκινο.

Στον Αλγόριθμο 3.5 περιγράφεται ένας αποδοτικός αλγόριθμος που τηρεί τους περιορισμούς της στρατηγικής Blue Fringe.

Αλγόριθμος: Blue Fringe

Είσοδος: Ένα PTA

Έξοδος: Ένα ντετερμινιστικό SFA

αρχή

θέσε την αρχική κατάσταση του PTA κόκκινη·

θέσε τα παιδιά της αρχικής κατάστασης μπλε·

θέσε τις υπόλοιπες καταστάσεις άσπρες·

όσο υπάρχει μπλε κατάσταση

βαθμολόγησε όλες τις συγχωνεύσεις κόκκινων/μπλε·

εάν υπάρχουν μπλε καταστάσεις ασύμβατες με όλες τις κόκκινες

μετάτρεψε την κοντινότερη στην αρχική από αυτές σε κόκκινη·

θέσε μπλε τα άσπρα παιδιά της·

αλλιώς

συγχώνευσε το ζευγάρι κόκκινης/μπλε κατάστασης με τη μεγαλύτερη βαθμολογία·

θέσε μπλε τα άσπρα παιδιά της νέας κατάστασης·

τέλος

Αλγόριθμος 3.5 Ο αλγόριθμος Blue Fringe

Ο αλγόριθμος αυτός μπορεί να προσαρμοστεί κατάλληλα για εκμάθηση πιθανοτικής κανονικής γραμματικής από ένα σύνολο θετικών παραδειγμάτων S^+. Σε αυτή την περίπτωση ο αλγόριθμος ξεκινά με ένα PPTA. Για τη βαθμολόγηση των συγχωνεύσεων μπορεί να χρησιμοποιηθεί οποιοσδήποτε στατιστικός έλεγχος. Παρακάτω παρουσιάζεται μία προσαρμογή του μετρικού που βασίζεται στο φράγμα Hoeffding και χρησιμοποιείται από τον Alergia.

Υπενθυμίζεται ότι δύο καταστάσεις θεωρούνται διαφορετικές όταν ισχύει η ανίσωση:

$$\left|\frac{f}{n} - \frac{f'}{n'}\right| > \sqrt{\frac{1}{2}\log\frac{2}{\alpha}}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n'}}\right)$$

όπου n, n' το πλήθος των συμβολοσειρών που φτάνουν στους δύο κόμβους, f, f' το πλήθος των συμβολοσειρών που τερματίζουν στον αντίστοιχο κόμβο ή ακολουθούν μία συγκεκριμένη μετάβαση και α μία παράμετρος.

Επιλύοντας ως προς α έχουμε:

$$2\,\mathrm{e}^{-2k} < \alpha \quad \text{όπου}$$

$$k = \left(\frac{f \cdot n' - f' \cdot n}{n'\sqrt{n} + n\sqrt{n'}}\right)$$

Το αριστερό μέλος της ανίσωσης λέγεται *p*-τιμή (*p*-value) και μπορεί να ιδωθεί ως η μικρότερη τιμή του α για την οποία ο στατιστικός έλεγχος απορρίπτει την υπόθεση ότι οι δύο καταστάσεις είναι όμοιες ως προς συγκεκριμένη μετάβαση. Δηλαδή η *p*-τιμή εκφράζει την ομοιότητα των δύο καταστάσεων. Η *p*-τιμή, όπως έχει οριστεί, παίρνει τιμές στο διάστημα [0,2].

Κάνοντας χρήση της παραπάνω έννοιας θεωρούμε ως μέτρο βαθμολόγησης δύο καταστάσεων την ελάχιστη τιμή των *p*-τιμών για όλους τους ελέγχους των μεταβάσεων καθώς και για τον έλεγχο της πιθανότητας να είναι οι καταστάσεις τερματικές. Αν η τιμή αυτή είναι μικρότερη ή ίση του α, τότε οι δύο καταστάσεις δε θεωρούνται συγχωνεύσιμες, ενώ όσο μεγαλύτερη είναι η τιμή τόσο πιο συμβατές θεωρούνται οι καταστάσεις. Η παρακάτω μέθοδος βαθμολόγησης της συγχώνευσης (Αλγόριθμος 3.6) αποτελεί τροποποίηση της μεθόδου ελέγχου συμβατότητας του Alergia και χρησιμοποιεί στην πράξη τη διαφορά της *p*-τιμής από το α ως τιμή βαθμολόγησης.

Η συγχώνευση (Αλγόριθμος 3.7) του ζευγαριού καταστάσεων με τη μεγαλύτερη βαθμολογία γίνεται με παρόμοιο τρόπο όπως και στον αλγόριθμο Alergia. Ειδική μέριμνα πρέπει να ληφθεί όσον αφορά τα χρώματα των καταστάσεων. Συγκεκριμένα η νέα κατάσταση θεωρείται κόκκινη, ενώ στην περίπτωση που στην αναδρομή συγχωνεύεται ένας κόκκινος με έναν άσπρο κόμβο, απαιτείται ο χρωματισμός των άσπρων παιδιών του νέου κόμβου σε μπλε.

Ο αλγόριθμος που βασίζεται στη στρατηγική Blue Fringe και παρουσιάστηκε παραπάνω έχει άνω όριο στη χρονική πολυπλοκότητα $P \cdot H^3$, όπου P είναι το μέγεθος του αρχικού PTA και H το πλήθος των κόμβων στην τελική υπόθεση.

Αλγόριθμος: mergeScore(q_i, q_j, α)

Είσοδος: q_i, q_j : καταστάσεις

α : παράμετρος

Έξοδος: Πραγματική τιμή

```
begin
    if pvalue(C(q_i), C(q_j), C(q_i,#), C(q_j,#)) < α
        return 0
    for each a∈Σ
        if pvalue(C(q_i), C(q_j), C(q_i,a), C(q_j,a)) < α
            return 0
        if mergeScore(δ(q_i,a), δ(q_j,a), α) <= 0
            return 0
    return min(pvalue) - α
end
```

Αλγόριθμος 3.6 Βαθμολόγηση συγχωνεύσεων στον Blue Fringe

Αλγόριθμος: merge(A, q_i, q_j)

Είσοδος: A : το PPTA

q_i, q_j : καταστάσεις προς συγχώνευση

```
begin
    A = A - {q_i,q_j}
    A = A ∪ q'
    q'.color = RED
    if q_i is RED and q_j is WHITE
        χρωμάτισε τα άσπρα παιδιά της q' μπλε
    C(q') = C(q_i) + C(q_j)
    C(q',#) = C(q_i,#) + C(q_j,#)
    for each a ∈ Σ
        C(q',a) = C(q_i,a) + C(q_j,a)
        merge(A, δ(q_i,a), δ(q_j,a))
end
```

Αλγόριθμος 3.7 Διαδικασία συγχώνευσης στον Blue Fringe

4

Μέθοδοι Μοντελοποίησης της Πλοήγησης των Χρηστών

Στην πρώτη ενότητα παρουσιάζονται κάποιες προσεγγίσεις στο αντικείμενο της Εξόρυξης Προτύπων Πλοήγησης. Στη δεύτερη ενότητα παρουσιάζεται αναλυτικά η μέθοδος μοντελοποίησης της πλοήγησης των χρηστών στον Παγκόσμιο Ιστό που προτείνεται στην παρούσα εργασία.

4.1 Σχετικές Εργασίες

Στο πλαίσιο της Εξόρυξης Γνώσης από Δεδομένα (βλ. Ενότητα 2.1) έχουν αναπτυχθεί στο πρόσφατο παρελθόν διάφορες τεχνικές μοντελοποίησης της πλοήγησης των χρηστών σε ένα ιστοχώρο αξιοποιώντας δεδομένα χρήσης του Παγκόσμιου Ιστού. Στόχος ενός τέτοιου μοντέλου είναι συνήθως η δυνατότητα πρότασης σχετικών συνδέσμων στο χρήστη ή εν γένει η πρόβλεψη των επόμενων σελίδων που θα ζητηθούν από αυτόν. Το νέο αυτό αντικείμενο ονομάζεται Εξόρυξη Προτύπων Πλοήγησης (Navigation Pattern Discovery) και είναι στην ουσία εφαρμογή της Εξόρυξης Προτύπων Διαδοχής στο πεδίο της Εξατομίκευσης υπηρεσιών στον Ιστό. Όπως αναφέρθηκε και στο Κεφ. 2 οι τεχνικές της Εξόρυξης Προτύπων Διαδοχής χωρίζονται σε ντετερμινιστικές, που καταγράφουν τη συμπεριφορά των χρηστών ως προς την πλοήγηση στον Ιστό, και στοχαστικές, που αξιοποιούν την ακολουθία των επισκεφθέντων σελίδων για την πρόβλεψη επόμενων επισκέψεων.

4.1.1 Ντετερμινιστικές Προσεγγίσεις

Ένα παράδειγμα ντετερμινιστικής τεχνικής δίνεται στο [SFW99], όπου χρησιμοποιείται το εργαλείο WUM (Web Utilization Miner) για ανακάλυψη προτύπων. Η μονάδα επεξεργασίας της γλώσσας εξόρυξης του WUM, της MINT, εξάγει κανόνες διαδοχής από προεπεξεργασμένα δεδομένα χρήσης του Ιστού. Η γλώσσα αυτή υποστηρίζει κατηγορήματα, με τα οποία μπορεί να προσδιοριστεί το περιεχόμενο, η δομή και στατιστικά στοιχεία των προτύπων πλοήγησης. Ο επεξεργαστής της MINT παρέχει διαδραστική εξόρυξη και χρησιμοποιεί περιορισμούς που καθορίζονται από κάποιον ειδικό. Πρόκειται δηλαδή για μία ημιαυτόματη διαδικασία, γεγονός που αποτελεί το μειονέκτημα του συστήματος.

Μία εναλλακτική μέθοδος προτείνεται στο [PPK+00]. Σύμφωνα με αυτή, οι σύνοδοι χρήσης αναπαρίστανται από τις μεταβάσεις μεταξύ των σελίδων της κάθε συνόδου. Έπειτα με ομαδοποίηση των δεδομένων αυτών παράγονται οι λεγόμενες κοινότητες χρηστών ως προς τη συμπεριφορά τους κατά την πλοήγηση. Αν και η μέθοδος αυτή παρέχει περιορισμένη μοντελοποίηση των προτύπων, η εμπειρική της αξιολόγηση δείχνει ότι μπορεί να παραγάγει ενδιαφέροντα πρότυπα πλοήγησης.

Μία άλλη ντετερμινιστική προσέγγιση χρησιμοποιείται από το εργαλείο Clementine της SPSS. Αυτό χρησιμοποιεί τον αλγόριθμο ανακάλυψης προτύπων διαδοχής CAPRI (Clementine A-Priori Intervals), ο οποίος εκτός της ανακάλυψης συσχετίσεων μεταξύ αντικειμένων (π.χ. ιστοσελίδες) βρίσκει επίσης τη σειρά με την οποία τα αντικείμενα έχουν προσπελαστεί αξιοποιώντας χρονική πληροφορία. Μία διαδικασία ανακάλυψης κατά CAPRI περιλαμβάνει τρεις φάσεις: στην πρώτη, την a priori, εντοπίζονται τα συχνά εμφανιζόμενα σύνολα αντικειμένων· κατόπιν, στη φάση ανακάλυψης σχηματίζονται δέντρα ακολουθιών, ένα για κάθε πιθανό αρχικό αντικείμενο· τέλος, στη φάση αποκοπής εξαλείφονται οι επικαλυπτόμενες ακολουθίες. Ο CAPRI μπορεί να αποκαλύψει κοινές ακολουθίες υπό κάποιους περιορισμούς που τίθενται από το χρήστη.

4.1.2 Στοχαστικές Προσεγγίσεις

Οι περισσότερες στοχαστικές μέθοδοι Εξόρυξης Προτύπων Διαδοχής κάνουν χρήση μαρκοβιανών μοντέλων για την πρόβλεψη του συνδέσμου (link) που θα επιλέξει ο χρήστης, λόγω της καταλληλότητάς τους στη μοντελοποίηση ακολουθιακών διαδικασιών. Μία από τις πρώτες εφαρμογές των μαρκοβιανών μοντέλων στα δεδομένα χρήσης του Ιστού παρουσιάζεται στο [Bes95]. Εκεί χρησιμοποιείται ένα κρυφό μαρκοβιανό μοντέλο (hidden Markov model) πρώτης τάξης για την πρόβλεψη του επόμενου συνδέσμου που ενδέχεται να ακολουθήσει ο χρήστης μέσα σε δεδομένο χρονικό διάστημα. Στο [Sar00] χρησιμοποιούνται μαρκοβιανές αλυσίδες για τη μοντελοποίηση της διαδοχής ιστοσελίδων. Παρόμοια

προσέγγιση ακολουθείται στο [Zhu01], όπου επιπλέον αξιοποιείται πληροφορία για τη σελίδα από την οποία έγινε η μετάβαση στην τρέχουσα. Στο [CHM+00] παρουσιάζεται μια μέθοδος που χρησιμοποιεί ένα μίγμα από μαρκοβιανά μοντέλα, ένα για κάθε ομάδα χρηστών, με σκοπό το χαρακτηρισμό της συμπεριφοράς τους ως προς την πλοήγηση στον Ιστό.

Μία διαφορετική τεχνική χρησιμοποιεί τέσσερα διαφορετικά μαρκοβιανά μοντέλα για την πρόβλεψη ιστοσελίδων μέσα σε μια υβριδική δομή που λέγεται maxHybrid [AZN99]. Τα τέσσερα μοντέλα είναι: το χρονικό μαρκοβιανό μοντέλο, που προβλέπει τον επόμενο σύνδεσμο βασισμένο μόνο στην τελευταία σελίδα που ζητήθηκε, το μαρκοβιανό μοντέλο δεύτερης τάξης, που βασίζει την πρόβλεψη του συνδέσμου στις δύο τελευταίες σελίδες που ζητήθηκαν, το χωρικό μαρκοβιανό μοντέλο, που χρησιμοποιεί τη σελίδα από την οποία έγινε η μετάβαση, και το συνδεδεμένο χωροχρονικό μαρκοβιανό μοντέλο, που συνδυάζει τις δύο πρακτικές. Όταν μία σελίδα ζητηθεί, τα τέσσερα μοντέλα υπολογίζουν την πιθανότητα της επόμενης σελίδας που θα ζητηθεί και ο maxHybrid επιλέγει αυτό με τη μεγαλύτερη πιθανότητα ορθής πρόβλεψης.

Στο [PP99] ακολουθείται μια άλλη προσέγγιση, η οποία εξάγει τις μακρύτερες ακολουθίες που έχουν συχνότητα πάνω από ένα κατώφλι. Αυτές λέγονται Μακρύτερες Επαναλαμβανόμενες Υπακολουθίες (Longest Repeating Subsequences - LRS) και χρησιμοποιούνται ως είσοδος σε δύο ειδών μαρκοβιανά μοντέλα για την πρόβλεψη των επόμενων αιτήσεων. Το ένα μοντέλο είναι όμοιο με ένα μαρκοβιανό μοντέλο πρώτης τάξης και το άλλο με ένα μοντέλο k-τάξης. Η χρήση των αιτήσεων LRS οδηγεί σε μείωση της υπολογιστικής πολυπλοκότητας, η οποία είναι γενικά σημαντικό πρόβλημα στα μοντέλα υψηλής τάξης.

Στο [BL99] παρουσιάζεται μία άλλη στοχαστική προσέγγιση στην Εξόρυξη Προτύπων Διαδοχής από συνόδους χρήσης, οι οποίες μοντελοποιούνται με μία πιθανοτική γραμματική υπερκειμένου (hypertext probabilistic grammar - HPG). Αυτή είναι μία ειδική περίπτωση κανονικής γραμματικής που γενικά επιχειρεί τη μοντελοποίηση δομών υπερκειμένου, όπως είναι ο Παγκόσμιος Ιστός. Στο πλαίσιο αυτό, οι ιστοσελίδες αναπαρίστανται ως μη τερματικά σύμβολα της γραμματικής (που επίσης ισοδυναμούν εδώ με τερματικά), οι σύνδεσμοι ανάμεσα στις ιστοσελίδες ως κανόνες παραγωγής και οι ακολουθίες ιστοσελίδων ως συμβολοσειρές της αντίστοιχης γλώσσας. Ένας αλγόριθμος αναζήτησης κατά πλάτος στον κατευθυνόμενο γράφο που αναπαριστά τη γραμματική χρησιμοποιείται για την αναγνώριση συμβολοσειρών που μπορούν να περιγράψουν τη συμπεριφορά των χρηστών ως προς την πλοήγηση του ιστοχώρου. Ο αλγόριθμος είναι γενικά αποδοτικός, αλλά η έξοδός του εξαρτάται σημαντικά από τις αρχικές παραμέτρους.

Στην πτυχιακή του εργασία [Kar03] ο Καραμπατζιάκης χρησιμοποιεί μεθόδους Συμπερασμού Γραμματικών για τη μοντελοποίηση της πλοήγησης των χρηστών σε έναν ιστοχώρο.

Θεωρώντας τις ιστοσελίδες ως τερματικά σύμβολα μίας πιθανοτικής κανονικής γλώσσας και τις ακολουθίες ιστοσελίδων ως συμβολοσειρές της αντίστοιχης γλώσσας, κατασκευάζεται αρχικά ένα PPTA ή ένα πιθανοτικό αυτόματο υπερκειμένου (HPA) από τα δεδομένα χρήσης του Ιστού που έχουν συλλεχθεί από τον εξυπηρετητή ενός ιστοχώρου. Στο αρχικό αυτόματο εφαρμόζεται στη συνέχεια ο αλγόριθμος Συμπερασμού Γραμματικών Alergia ή ο Blue Fringe. Οι αλγόριθμοι αυτοί ελέγχουν τη συμβατότητα των καταστάσεων του αυτομάτου ως προς τις πιθανότητες των μεταβάσεων και επιτελούν συγχωνεύσεις καταστάσεων. Με αυτό τον τρόπο επάγεται από το αρχικό αυτόματο το αυτόματο-στόχος. Ο τελικός γράφος μπορεί να χρησιμοποιηθεί για να προταθούν σελίδες στον επισκέπτη του ιστοχώρου.

Τέλος στο [PP05] παρουσιάζεται μία διαφορετική προσέγγιση, οι Κατάλογοι Ιστού για Κοινότητες (Community Web Directories), με στόχο την αντιμετώπιση του προβλήματος του υπερβολικού όγκου πληροφοριών στον Ιστό. Εδώ οι κατάλογοι Ιστού θεωρούνται ιεραρχίες εννοιών και η εξατομίκευση πραγματοποιείται με την κατασκευή μοντέλων κοινοτήτων αξιοποιώντας δεδομένα χρήσης από αρχεία καταγραφής ενός ISP. Για το σκοπό αυτό εφαρμόζεται μία νέα μέθοδος Εξόρυξης Γνώσης από Δεδομένα, που λέγεται Community Directory Miner. Για τη μοντελοποίηση χρησιμοποιείται επίσης Πιθανοτική Ανάλυση Λανθάνουσας Σημασιολογίας (Probabilistic Latent Semantic Analysis).

4.2 Νέα Μέθοδος Κατασκευής Μοντέλου

Η πλειοψηφία των εργασιών που αναφέρονται στην προηγούμενη ενότητα αφορούν στην Εξόρυξη Προτύπων Πλοήγησης από δεδομένα χρήσης ενός μόνο ιστοχώρου με σκοπό την παροχή εξατομικευμένων υπηρεσιών για τον ιστοχώρο αυτό. Στην παρούσα εργασία αντίθετα επιχειρείται η μοντελοποίηση της πλοήγησης των χρηστών σε ολόκληρο τον Παγκόσμιο Ιστό. Το πρόβλημα αυτό καθίσταται όμως δυσκολότερο λόγω του μεγάλου όγκου και της ανομοιογένειας του Παγκόσμιου Ιστού, σε σύγκριση με ένα συγκεκριμένο ιστοχώρο.

Στην ενότητα αυτή παρουσιάζεται στο πλαίσιο του Συμπερασμού Γραμματικών μία μέθοδος Εξόρυξης Προτύπων Πλοήγησης από δεδομένα χρήσης του Παγκόσμιου Ιστού, με το όνομα CANUMGI (Content-Aware Navigational User Modeling with Grammatical Inference). Πρόκειται για επέκταση των αλγορίθμων Alergia και Blue Fringe στην κατεύθυνση της Ανάκτησης Πληροφοριών. Η μέθοδος εκτελείται off-line και κατασκευάζει το μοντέλο, που είναι ένα στοχαστικό πεπερασμένο αυτόματο (SFA). Το αυτόματο αυτό μπορεί έπειτα να χρησιμοποιηθεί on-line για την πρόταση συνδέσμων σελίδων (links) σε χρήστες.

Για την κατασκευή του μοντέλου ακολουθείται η παρακάτω διαδικασία. Ως είσοδος θεωρούνται δεδομένα χρήσης του Παγκόσμιου Ιστού που μπορούν να ληφθούν από αρχεία καταγραφής μιας εταιρείας ISP. Τα δεδομένα αυτά αποτελούνται από ακολουθίες

ιστοσελίδων που επισκέφτηκαν χρήστες-πελάτες της εταιρείας ISP μέσα σε κάποιο χρονικό διάστημα. Οι σελίδες αυτές είναι χωρισμένες σε συνόδους χρήσης. Όπως και στην εργασία [Kar03], οι ιστοσελίδες θεωρούνται τερματικά σύμβολα μίας πιθανοτικής κανονικής γραμματικής. Επίσης μια ακολουθία ιστοσελίδων (δηλαδή μια σύνοδος χρήσης) θεωρείται συμβολοσειρά της αντίστοιχης γλώσσας.

Αρχικά κατασκευάζεται από τα δεδομένα χρήσης ένα πιθανοτικό δενδρικό αυτόματο προθημάτων (PPTA), σαν αυτό του Σχήματος 3.3, στο οποίο τα γράμματα του λατινικού αλφαβήτου συμβολίζουν σελίδες. Η κατασκευή του PPTA γίνεται όπως περιγράφεται στην Ενότητα 3.4 και με σκοπό την εφαρμογή σε αυτό μίας μεθόδου Συμπερασμού Γραμματικών, όπως είναι οι αλγόριθμοι Alergia και Blue Fringe. Ωστόσο, η εφαρμογή των αλγορίθμων αυτών στην κλασική τους εκδοχή δεν είναι αποτελεσματική για τον παρακάτω λόγο. Εξαιτίας του μεγάλου όγκου του Παγκόσμιο Ιστού, τα δεδομένα χρήσης είναι εξαιρετικά ανομοιογενή. Αυτό έχει ως αποτέλεσμα οι περισσότερες καταστάσεις του PPTA να προκύπτουν ασύμβατες, καθώς η συμβατότητα δύο καταστάσεων καθορίζεται από την ομοιότητα των μεταβάσεων, και συνεπώς να μη γίνονται αρκετές συγχωνεύσεις καταστάσεων.

4.2.1 Συμβατότητα με βάση το περιεχόμενο

Για να αντιμετωπιστεί αυτό το πρόβλημα απαιτείται η εισαγωγή ενός νέου μέτρου συμβατότητας των καταστάσεων που να λαμβάνει υπόψη του την ομοιότητα των ιστοσελίδων ως προς το περιεχόμενό τους. Για την εφαρμογή αυτού του μέτρου πρέπει να διατίθεται επιπλέον αρχική πληροφορία που να περιγράφει το περιεχόμενο των σελίδων. Στην παρούσα εργασία ακολουθείται η παρακάτω προσέγγιση, που προέρχεται από το πεδίο της Ανάκτησης Πληροφοριών. Κάθε σελίδα χαρακτηρίζεται από ένα σύνολο λέξεων-κλειδιών (keywords), που έχουν εξαχθεί από τις σελίδες στη φάση της προεπεξεργασίας. Αν διατάξουμε το σύνολο των λέξεων-κλειδιών όλων των σελίδων, τότε μία σελίδα χαρακτηρίζεται από το διάνυσμα $x = (x_1, x_2, ..., x_i, ...)$, όπου το x_i παίρνει την τιμή 1, αν η σελίδα περιέχει τη λέξη-κλειδί i, αλλιώς 0. Σημειώνεται ότι το διάνυσμα αυτό είναι αραιό, καθώς σε κάθε σελίδα αντιστοιχούν μόνο λίγες από το σύνολο των λέξεων-κλειδιών. Ως μέτρο της ομοιότητας των σελίδων χρησιμοποιείται το μετρικό του συνημιτόνου που παρουσιάστηκε στην Υποενότητα 2.3.2 και επαναλαμβάνεται εδώ:

$$COSINE(x, y) = \frac{\sum_i x_i y_i}{\sqrt{\left(\sum_i x_i^2\right)\left(\sum_i y_i^2\right)}}$$

Στον παραπάνω τύπο τα x και y είναι τα διανύσματα λέξεων-κλειδιών των προς έλεγχο σελίδων. Το μετρικό αυτό είναι κατάλληλο για τη συγκεκριμένη εφαρμογή, διότι λαμβάνει

υπόψη του μόνο τις μη μηδενικές τιμές των διανυσμάτων, κάτι που είναι σημαντικό όταν τα δεδομένα είναι αραιά.

Στο πλαίσιο του Συμπερασμού Γραμματικών πρέπει να επεκτείνουμε την έννοια του διανύσματος, καθώς εδώ ασχολούμαστε με καταστάσεις ενός αυτομάτου, οι οποίες μπορεί εν γένει να περιέχουν μία ή περισσότερες σελίδες ύστερα από μία σειρά συγχωνεύσεων. Πρόκειται δηλαδή για ομάδες (clusters) από σελίδες. Έτσι μία κατάσταση που περιέχει k σελίδες χαρακτηρίζεται από το διάνυσμα $x=(x_1,x_2,...,x_i,...)$, όπου το x_i παίρνει πραγματικές τιμές στο διάστημα $[0,1]$ και εκφράζει το ποσοστό των σελίδων της κατάστασης που έχουν τη λέξη-κλειδί i. Αν θεωρήσουμε ότι κάθε σελίδα μιας ομάδας αντιπροσωπεύεται από το διάνυσμά της στο χώρο με διαστάσεις τις λέξεις-κλειδιά, τότε μπορούμε να θεωρήσουμε το διάνυσμα x ως ένα σημείο που αντιστοιχεί στο κέντρο βάρους της ομάδας και την εκπροσωπεί. Μπορούμε λοιπόν να ορίσουμε ως μέτρο ομοιότητας περιεχομένου δύο καταστάσεων q_i, q_j, similarity(q_i,q_j), το συνημίτονο των αντίστοιχων διανυσμάτων.

4.2.2 Η Μέθοδος CANUMGI-A

Η πρώτη παραλλαγή της μεθόδου CANUMGI, που προτείνεται σε αυτή τη διπλωματική, αποτελεί επέκταση του αλγορίθμου Συμπερασμού Γραμματικών Alergia. Η βασική διαδικασία (Αλγόριθμος. 3.1) παραμένει η ίδια· ο αλγόριθμος ξεκινά από το PPTA και ελέγχει τους κόμβους ως προς τη συμβατότητα. Κατά τον έλεγχο της συμβατότητας συνδυάζεται το παραπάνω μετρικό περιεχομένου με το μετρικό χρήσης που μεταχειρίζεται ο κλασικός Alergia. Στη συνέχεια (Αλγόριθμος 4.1) παρουσιάζεται ο αλγόριθμος ελέγχου συμβατότητας της μεθόδου.

Στον αλγόριθμο αυτό υπολογίζονται οι *p*-τιμές για όλες τις μεταβάσεις καθώς και για την πιθανότητα οι καταστάσεις να είναι τερματικές, όπως και στον κλασικό Alergia. Επιπλέον υπολογίζεται το μετρικό περιεχομένου ανάμεσα στις δύο καταστάσεις. Λέμε ότι δύο καταστάσεις είναι συμβατές ως προς ένα μετρικό, όταν η τιμή του είναι πάνω από το αντίστοιχο κατώφλι που έχουμε θέσει. Για τον προσδιορισμό της (συνολικής) συμβατότητας δύο καταστάσεων ελέγχουμε τη συμβατότητα ως προς τα δύο μετρικά και στη συνέχεια λαμβάνουμε τη διάζευξη ή τη σύζευξη των δύο αποτελεσμάτων. Επειδή όμως οι *p*-τιμές είναι περισσότερες από μία, η τιμή του μετρικού χρήσης προκύπτει ως συνάρτηση *g* αυτών των *p*-τιμών. Στο σημείο αυτό προτείνονται δύο τέτοιες συναρτήσεις *g*:

- min(p-values), η ελάχιστη τιμή των *p*-τιμών
- average(p-values), η μέση τιμή των *p*-τιμών

Η πρώτη περίπτωση είναι ισοδύναμη με τον έλεγχο στον κλασικό Alergia, όπου απαιτείται να ισχύει η ομοιότητα ως προς όλες τις μεταβάσεις καθώς και ως προς την πιθανότητα

τερματισμού. Επομένως η σύζευξη του μετρικού χρήσης σε αυτή την περίπτωση με το μετρικό περιεχομένου δεν έχει νόημα, διότι έτσι ο έλεγχος γίνεται ακόμα πιο αυστηρός. Αντίθετα η διάζευξη των μετρικών συμβάλλει στη χαλάρωση της αυστηρότητας του ελέγχου, καθώς προσφέρει ένα εναλλακτικό κριτήριο.

Αλγόριθμος: compatible(q_i, q_j, α)

Είσοδος: q_i, q_j : καταστάσεις

α : κατώφλι μετρικού χρήσης

θ : κατώφλι μετρικού περιεχομένου

Έξοδος: Αληθές, αν οι δύο καταστάσεις είναι συμβατές

```
begin
    calculate similarity(q_i , q_j)
    calculate pvalue(C(q_i), C(q_j), C(q_i,#), C(q_j,#))
    foreach a∈Σ
        calculate pvalue(C(q_i), C(q_j), C(q_i,a), C(q_j,a))
        if not compatible(δ(q_i,a), δ(q_j,a), α)
            return false
    if similarity > θ or/and g(pvalues) > α
        return true
    return false
end
```

Αλγόριθμος 4.1 Έλεγχος συμβατότητας της μεθόδου CANUMGI-A

Στη δεύτερη περίπτωση το μετρικό χρήσης γίνεται πιο χαλαρό, καθώς η τιμή του δεν εξαρτάται απόλυτα από μία ελάχιστη τιμή. Έτσι είναι ανεκτικό σε περιπτώσεις που δύο καταστάσεις είναι όμοιες ως προς τους περισσότερους ελέγχους αλλά όχι ως προς μερικούς. Εδώ μπορεί πλέον το μετρικό χρήσης να συνδυασθεί με το μετρικό περιεχομένου τόσο συζευκτικά όσο και διαζευκτικά. Πάντως στην περίπτωση της διάζευξης, ο έλεγχος αναμένεται να είναι εξαιρετικά χαλαρός.

Τέλος πρέπει να σημειωθεί ότι, όπως και στον κλασικό Alergia, γίνεται κατά τον έλεγχο συμβατότητας αναδρομικός έλεγχος των καταστάσεων όπου καταλήγουν οι μεταβάσεις για την εξασφάλιση του ντετερμινισμού του αυτομάτου.

Όσον αφορά το μετρικό περιεχομένου σημειώνεται ότι για πρακτικούς λόγους σε κάθε κόμβο τηρείται όχι το διάνυσμα πραγματικών τιμών x όπως ορίστηκε στην προηγούμενη υποενότητα, αλλά ένα διάνυσμα v ακέραιων τιμών που εκφράζουν το πόσες σελίδες του

κόμβου (της ομάδας) έχουν την αντίστοιχη λέξη-κλειδί. Επίσης χωριστά τηρείται το πλήθος k των σελίδων που περιέχει ένας κόμβος. Επομένως ισχύει $x = v/k$.

Με τη συγχώνευση των δύο καταστάσεων, ο νέος κόμβος αποτελεί μια ομάδα (cluster), που περιέχει τόσο τις σελίδες που προέρχονται από τον ένα συγχωνευμένο κόμβο όσο και από τον άλλο. Μία ομάδα ενδέχεται να περιέχει περισσότερες από μία φορές την ίδια σελίδα, αν αυτή περιεχόταν και στις δύο συνιστώσες ομάδες. Αυτό είναι επιθυμητό, διότι μία σελίδα που εμφανίζεται πολλές φορές είναι πιθανόν σημαντικότερη και πρέπει συνεπώς να λαμβάνεται περισσότερο υπόψη στον χαρακτηρισμό της ομάδας ως προς το περιεχόμενο με βάση το διάνυσμα λέξεων-κλειδιών. Κατά τη συγχώνευση (Αλγόριθμος 4.2), το διάνυσμα (ακέραιων τιμών) του νέου κόμβου τίθεται ίσο με το άθροισμα των διανυσμάτων των δύο κόμβων από τους οποίους αυτός προέκυψε και το πλήθος των σελίδων k επίσης ίσο με το άθροισμα των σελίδων των δύο κόμβων.

Αλγόριθμος: merge(A, q_i , q_j)

Είσοδος: A : το PPTA

q_i , q_j : καταστάσεις προς συγχώνευση

begin

A = A − $\{q_i, q_j\}$

A = A ∪ q′

ID of q′ = min(IDs of q_i, q_j)

vector(q′) = vector(q_i) + vector(q_j)

k′ = k_i + k_j

$C(q') = C(q_i) + C(q_j)$

$C(q', \#) = C(q_i, \#) + C(q_j, \#)$

for each a ∈ Σ

C(q′, a) = C(q_i, a) + C(q_j, a)

merge(A, δ(q_i, a), δ(q_j, a))

end

Αλγόριθμος 4.2 Διαδικασία συγχώνευσης της μεθόδου CANUMGI-A

Ωστόσο ο τρόπος με τον οποίο ο Alergia επιλέγει τις συγχωνεύσεις δεν είναι ικανοποιητικός. Συγκεκριμένα, εκτελώντας τις επαναλήψεις στους δύο βρόχους, επιλέγει να συγχωνεύσει το πρώτο συμβατό ζευγάρι που θα εντοπίσει, χωρίς να ερευνά την ύπαρξη πιθανώς πιο συμβατών ζευγαριών. Επιπλέον ένας συγχωνευμένος κόμβος παίρνει ένα χαμηλό αύξοντα αριθμό με αποτέλεσμα να προηγείται κατά τη διαδικασία σύγκρισης ενός νέου κόμβου (που περιέχει μία μόνο σελίδα) με τους υπόλοιπους. Με αυτό τον τρόπο ο αλγόριθμος τείνει να

ευνοεί τη δημιουργία ενός μεγάλου κόμβου που να περιέχει την πλειοψηφία των σελίδων, καθώς το μετρικό περιεχομένου ανάμεσα σε μία ομάδα σελίδων και σε μία σελίδα έχει συνήθως μία μη αμελητέα τιμή. Για το λόγο αυτό αναζητήθηκε μία μέθοδος που να επιλέγει με πιο αντικειμενικό τρόπο τα προς συγχώνευση ζευγάρια.

4.2.3 Η Μέθοδος CANUMGI-B

Η δεύτερη παραλλαγή της μεθόδου CANUMGI αποτελεί επέκταση του αλγορίθμου Συμπερασμού Γραμματικών Blue Fringe. Ο αλγόριθμος αυτός επιλέχθηκε, διότι διαλέγει με πιο αποδοτικό τρόπο τις καταστάσεις που θα συγχωνευτούν, ακολουθώντας μία προσέγγιση άπληστης (greedy) αναζήτησης. Συγκεκριμένα επιλέγει σε κάθε βήμα από τα σύνολα κόκκινων και μπλε κόμβων το ζευγάρι εκείνο που έχει τη μεγαλύτερη αξία συγχώνευσης. Έτσι αναμένεται μια πιο ορθολογική κατανομή των σελίδων σε ομάδες. Από την άλλη ο Blue Fringe έχει το ενδεχομένως αρνητικό χαρακτηριστικό ότι από τη στιγμή που δύο καταστάσεις έχουν γίνει κόκκινες, δεν μπορούν πλέον να συγχωνευθούν μεταξύ τους, ακόμα και αν στην πορεία λόγω άλλων συγχωνεύσεων γίνουν και μεταξύ τους συμβατές.

Η μέθοδος χρησιμοποιεί τη βασική διαδικασία του Blue Fringe (Αλγόριθμος 3.5). Ο αλγόριθμος βαθμολόγησης της συμβατότητας δύο καταστάσεων παρουσιάζεται στον Αλγόριθμο 4.3.

Αλγόριθμος: mergeScore(q_i, q_j, α)

Είσοδος: q_i, q_j: καταστάσεις

α : κατώφλι μετρικού χρήσης

θ : κατώφλι μετρικού περιεχομένου

Έξοδος: Πραγματική τιμή

```
begin
    calculate similarity(q_i , q_j)
    calculate pvalue(C(q_i), C(q_j), C(q_i,#), C(q_j,#))
    for each a∈Σ
        calculate pvalue(C(q_i), C(q_j), C(q_i,a), C(q_j,a))
        if mergeScore(δ(q_i,a), δ(q_j,a), α) <= 0
            return 0
    return f(similarity-θ, g(pvalues)-α)
end
```

Αλγόριθμος 4.3 Βαθμολόγηση συγχωνεύσεων της μεθόδου CANUMGI-B

Ο αλγόριθμος βαθμολόγησης είναι παρόμοιος με τον αλγόριθμο ελέγχου συμβατότητας της μεθόδου CANUMGI-A, με τη διαφορά ότι ο νέος αλγόριθμος υπολογίζει μία αριθμητική αντί για λογική τιμή. Αυτή εκφράζει το πόσο συμβατές θεωρούνται οι δύο καταστάσεις. Θεωρούμε ότι θετική τιμή σημαίνει συμβατές καταστάσεις και ότι όσο μεγαλύτερη είναι αυτή η τιμή τόσο περισσότερο συμβατές θεωρούνται οι καταστάσεις.

Η τιμή που υπολογίζει ο αλγόριθμος μπορεί να είναι οποιαδήποτε συνάρτηση των δύο μετρικών. Στην παρούσα εργασία εξετάζονται τρεις συναρτήσεις βαθμολόγησης $f(a,b)$, όπου το a είναι η διαφορά της τιμής του μετρικού περιεχομένου από το αντίστοιχο κατώφλι και το b η διαφορά μίας συνάρτησης *g* των *p*-τιμών του μετρικού χρήσης από το αντίστοιχο κατώφλι:

- $\max(a,b)$, η μέγιστη τιμή των δύο μετρικών
- $\min(a,b)$, η ελάχιστη τιμή των δύο μετρικών
- $(1-w)\cdot a + w\cdot b$, όπου το $w \in [0,1]$ είναι παράμετρος βάρους

Η πρώτη περίπτωση αποτελεί την ποσοτικοποίηση του κριτηρίου συμβατότητας της CANUMGI-A που χρησιμοποιεί διάζευξη των δύο μετρικών. Αντίστοιχα η δεύτερη περίπτωση ισοδυναμεί με το κριτήριο που κάνει σύζευξη των δύο μετρικών. Η τρίτη περίπτωση αποτελεί ένα σταθμισμένο άθροισμα των δύο μετρικών. Αυτό προσφέρει έναν εύχρηστο τρόπο παραμετροποίησης του βάρους το οποίο προσδίδουμε στο κάθε μετρικό. Για όλες τις παραπάνω περιπτώσεις ως συνάρτηση *g* των *p*-τιμών του μετρικού χρήσης θεωρείται μία από τις δύο που παρουσιάστηκαν στο πλαίσιο της μεθόδου CANUMGI-A.

Είναι σημαντικό ότι, για να εφαρμοστούν οι παραπάνω συναρτήσεις *f*, πρέπει πρώτα να έχουν κανονικοποιηθεί οι τιμές από τα δύο μετρικά, έτσι ώστε να είναι συγκρίσιμες. Στο επόμενο κεφάλαιο περιγράφεται ο τρόπος με τον οποίο έγινε αυτή η κανονικοποίηση στα υπάρχοντα δεδομένα, καθώς εξαρτάται από τη μορφή των εκάστοτε δεδομένων.

Όσον αφορά τη συγχώνευση των καταστάσεων (Αλγόριθμος 4.4), αυτή γίνεται με τον τρόπο που παρουσιάστηκε στην Ενότητα 3.5 με τις απαραίτητες προσθήκες σχετικά με το μετρικό του περιεχομένου.

Αλγόριθμος: merge(A, q_i , q_j)

Είσοδος: A : το PPTA

q_i , q_j : καταστάσεις προς συγχώνευση

begin

A = A - $\{q_i, q_j\}$

A = A ∪ q′

vector(q′) = vector(q_i) + vector(q_j)

k′ = k_i + k_j

q′.color = RED

if q_i is RED **and** q_j is WHITE

χρωμάτισε τα άσπρα παιδιά της q′ μπλε

$C(q') = C(q_i) + C(q_j)$

$C(q', \#) = C(q_i, \#) + C(q_j, \#)$

for each a ∈ Σ

$C(q', a) = C(q_i, a) + C(q_j, a)$

merge(A, $\delta(q_i, a), \delta(q_j, a)$)

end

Αλγόριθμος 4.4 Διαδικασία συγχώνευσης της μεθόδου CANUMGI-B

4.2.4 Η Μέθοδος CANUMGI-C

Όπως έχει ήδη αναφερθεί, το γεγονός ότι τα δεδομένα χρήσης του Παγκόσμιου Ιστού χαρακτηρίζονται από μεγάλη ανομοιογένεια αποτελεί σημαντικό πρόβλημα της Εξόρυξης Προτύπων Πλοήγησης. Όντως παρατηρείται ότι μία σελίδα σπάνια υπάρχει πάνω από μία φορά στα δεδομένα χρήσης, γεγονός που καθιστά δύσκολη την εξόρυξη χρήσιμων προτύπων. Εδώ παρουσιάζεται μία τρίτη παραλλαγή της μεθόδου CANUMGI, που ακολουθεί μία νέα προσέγγιση για να αντιμετωπίσει το παραπάνω πρόβλημα. Σύμφωνα με αυτή, γίνεται αρχικά ομαδοποίηση (clustering) των σελίδων που υπάρχουν στα δεδομένα εκπαίδευσης σε έναν ορισμένο αριθμό ομάδων. Η ομαδοποίηση αυτή γίνεται με κριτήριο το περιεχόμενο των σελίδων. Από εδώ και στο εξής, κάθε σελίδα αντιπροσωπεύεται από την ομάδα στην οποία ανήκει. Έτσι, με τη μετάβαση από τις σελίδες στις ομάδες, επιτυγχάνεται μία μείωση του αριθμού των διαστάσεων (dimensionality reduction) του προβλήματος. Κίνητρο για τη διαδικασία αυτή αποτελεί η εκτίμηση ότι η σελίδες μπορούν να χωριστούν σε θεματικές κατηγορίες ως προς το περιεχόμενό τους και ότι οι σελίδες μιας κατηγορίας παρουσιάζουν ενδεχομένως ενιαία συμπεριφορά ως προς τα δεδομένα χρήσης. Αυτή η μείωση λοιπόν

ενδέχεται να αναδείξει την ύπαρξη κοινών προτύπων πλοήγησης για όμοιες σελίδες και να συμβάλει τελικά στην κατασκευή ενός πιο εκφραστικού μοντέλου. Στη συνέχεια περιγράφεται αναλυτικά η διαδικασία αυτή.

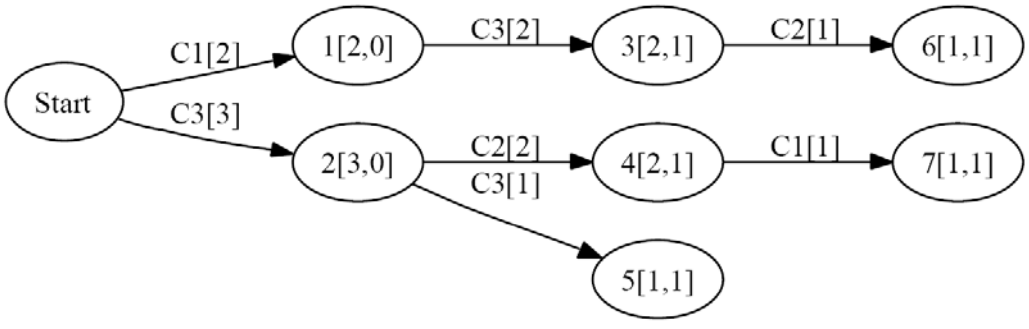
Σε πρώτη φάση γίνεται η ομαδοποίηση των σελίδων που υπάρχουν στις συνόδους χρήσης του δείγματος εκπαίδευσης. Οι σελίδες αναπαρίστανται, όπως και πριν, με το διάνυσμα των λέξεων-κλειδιών τους. Για την ομαδοποίηση χρησιμοποιείται ο αλγόριθμος k-means, που είναι μία μέθοδος διαμέρισης και παρουσιάστηκε στην Υποενότητα 2.3.3. Ο αλγόριθμος αυτός απαιτεί τον προσδιορισμό εκ των προτέρων του πλήθους των ομάδων που θα δημιουργηθούν.

Στη συνέχεια κατασκευάζεται το PPTA από τα δεδομένα χρήσης του δείγματος εκπαίδευσης. Οι μεταβάσεις του PPTA δεν αντιστοιχούν πλέον σε σελίδες αλλά στις ομάδες όπου αυτές ανήκουν. Αντίστοιχα, ένα κόμβος του αυτομάτου θεωρείται ότι περιέχει μία ολόκληρη ομάδα παρά μία σελίδα. Η ομάδα αυτή εκπροσωπείται από το διάνυσμα των λέξεων-κλειδιών, που έχει προκύψει από τα διανύσματα των σελίδων της, εκπροσωπείται δηλαδή από το κέντρο βάρους της. Όμοια με τον τρόπο που περιγράφεται στην CANUMGI-A, για κάθε ομάδα σελίδων τηρείται ένα διάνυσμα ακέραιων τιμών v, όπου καταγράφεται το πλήθος των σελίδων που έχουν την αντίστοιχη λέξη-κλειδί. Επίσης, τηρείται το συνολικό πλήθος των σελίδων της ομάδας k (ο πληθάριθμός της).

Έστω το παράδειγμα με το δείγμα $S^+ = \{\text{abc, de, ad, def, gb}\}$. Αν υποθέσουμε ότι οι σελίδες, που παριστάνονται με τα λατινικά γράμματα, έχουν ομαδοποιηθεί στις τρεις ομάδες $C1 = \{\text{a,f}\}$, $C2 = \{\text{c,e}\}$, $C3 = \{\text{b,d,g}\}$, τότε το δείγμα γίνεται

$$S^+ = \{C1C3C2, C3C2, C1C3, C3C2C1, C3C3\}$$

και το αντίστοιχο PPTA είναι:



Σχήμα 4.1 Το PPTA μετά τη μείωση διαστασιμότητας

Η κατασκευή του PPTA μπορεί επίσης να γίνει με έναν ελαφρώς διαφορετικό τρόπο. Συχνά παρατηρείται το φαινόμενο όλες οι σελίδες μιας ακολουθίας (υπακολουθίας μιας συνόδου χρήσης) να ανήκουν στην ίδια ομάδα. Αυτό αντιστοιχεί σε διαδοχικές μεταβάσεις στο PPTA για το ίδιο σύμβολο (την ίδια ομάδα), όπως οι μεταβάσεις $Start \to 2 \to 5$ στο παραπάνω

σχήμα. Εκτιμώντας ότι, αν περιορίσουμε τον αρχικό γράφο έτσι ώστε να περιλαμβάνει μόνο τις «πραγματικές» μεταβάσεις, δηλαδή τις μεταβάσεις που γίνονται από μια σελίδα κάποιας ομάδας σε μια σελίδα μιας άλλης ομάδας, τότε το μοντέλο θα είναι πιο εκφραστικό, μπορούμε να αγνοήσουμε τις διαδοχικές μεταβάσεις στην ίδια ομάδα (π.χ. τη μετάβαση $2 \to 5$ στο παραπάνω σχήμα). Με αυτό τον τρόπο ευνοείται περισσότερο η ανάδειξη των συσχετίσεων ανάμεσα στις ομάδες.

Μετά την κατασκευή του PPTA με κάποιον από τους δύο τρόπους που περιγράφηκαν, ακολουθεί η επαγωγή του αυτομάτου-στόχου, που μπορεί να γίνει με μία από τις μεθόδους CANUMGI-A και CANUMGI-B. Στα επόμενα θεωρούμε ότι χρησιμοποιείται η CANUMGI-B. Μία διαφοροποίηση υπάρχει ωστόσο στη θεώρηση του διανύσματος μίας κατάστασης. Όταν δύο κόμβοι του αυτομάτου συγχωνεύονται, τότε ο κόμβος που προκύπτει περιέχει όλες τις ομάδες σελίδων που περιείχαν οι αρχικοί κόμβοι. Αν όμως η ίδια ομάδα υπήρχε και στους δύο κόμβους, δε θα ήταν ορθό να την περιλάβουμε δύο φόρες στον τελικό κόμβο, όπως γίνεται στην περίπτωση που έχουμε απλώς σελίδες, διότι έτσι θα πριμοδοτείτο υπερβολικά μία ολόκληρη ομάδα, ακόμα κι αν αντιστοιχούσαν πρακτικά στον κόμβο λίγες μόνο σελίδες της. Έτσι, το διάνυσμα ενός νέου κόμβου δεν ισοδυναμεί πλέον με το άθροισμα των διανυσμάτων των συνιστώντων κόμβων, αλλά με το άθροισμα των διανυσμάτων των ομάδων που περιέχονται στο νέο αυτό κόμβο. Με παρόμοιο τρόπο προκύπτει και το πλήθος των σελίδων που περιέχονται σε έναν κόμβο. Αυτά φαίνονται στο επόμενο σχήμα (Αλγόριθμος 4.5), που παρουσιάζει τον αλγόριθμο συγχώνευσης καταστάσεων για την περίπτωση μείωσης διαστασιμότητας.

Αλγόριθμος: merge(A, q_i , q_j)

Είσοδος: A : το PPTA

q_i , q_j : καταστάσεις προς συγχώνευση

begin

A = A - $\{q_i, q_j\}$

A = A ∪ q′

vector(q′) = $\sum_{\text{clusters in q′}}$ vector

k′ = $\sum_{\text{clusters in q′}}$ k

q′.color = RED

if q_i is RED **and** q_j is WHITE

χρωμάτισε τα άσπρα παιδιά της q′ μπλε

$C(q') = C(q_i) + C(q_j)$

$C(q', \#) = C(q_i, \#) + C(q_j, \#)$

for each a ∈ Σ

$C(q', a) = C(q_i, a) + C(q_j, a)$

merge(A, $\delta(q_i, a), \delta(q_j, a)$)

end

Αλγόριθμος 4.5 Διαδικασία συγχώνευσης της μεθόδου CANUMGI-C

4.3 Χρήση του Νέου Μοντέλου για Εξατομικευμένη Πλοήγηση

Το αυτόματο που κατασκευάστηκε στην προηγούμενη ενότητα μπορεί να αξιοποιηθεί για την πρόταση σελίδων του Παγκόσμιου Ιστού σε ένα χρήστη. Η γενική ιδέα είναι ότι οι μεταβάσεις ενός χρήστη από σελίδα σε σελίδα αντιστοιχούν σε μεταβάσεις από κόμβο σε κόμβο στο γράφο. Έτσι, ευρισκόμενοι σε κάποιο κόμβο, επιχειρούμε να προβλέψουμε τις επόμενες σελίδες που θα επισκεφτεί ο χρήστης ή εναλλακτικά να του προτείνουμε πιθανώς ενδιαφέροντες συνδέσμους.

Η διαδικασία επιλογής σελίδων που θα προταθούν αναλύεται σε δύο στάδια. Κατά το πρώτο, προσδιορίζεται η κατάσταση (κόμβος του γράφου) στην οποία έχει μεταβεί το σύστημα ως προς τη μέχρι στιγμής πλοήγηση του χρήστη. Κατά το δεύτερο στάδιο, με βάση την κατάσταση αυτή επιλέγονται οι καλύτερες σελίδες από τις καταστάσεις-παιδιά της σύμφωνα με ορισμένα αξιολογικά κριτήρια που αναλύονται στη συνέχεια.

4.3.1 Εξατομικευμένη Πλοήγηση με τις Μεθόδους CANUMGI-A και CANUMGI-B

4.3.1.1 Διάσχιση του Γράφου

Ο τυπικός τρόπος με τον οποίο ένα ντετερμινιστικό αυτόματο αλλάζει κατάσταση είναι με την ανάγνωση ενός συμβόλου στην είσοδό του. Αναμένεται βέβαια ότι υπάρχει μία μετάβαση από την τρέχουσα κατάσταση για αυτό το σύμβολο. Για ένα αυτόματο όμως που επιχειρεί να περιγράψει την πλοήγηση των χρηστών στον Παγκόσμιο Ιστό, το σύνολο των πιθανών συμβόλων είναι το σύνολο των σελίδων του Παγκόσμιου Ιστού. Από τη στιγμή που τα δεδομένα με τα οποία εκπαιδεύτηκε το μοντέλο είναι πεπερασμένα, δε μπορούμε να αναμένουμε ότι θα αναπαρίσταται κάθε πιθανή μετάβαση στο γράφο. Επομένως είναι αναγκαίο να γίνει πιο χαλαρός ο τρόπος επιλογής της επόμενης κατάστασης. Στην παρούσα εργασία ακολουθήθηκε η προσέγγιση του Αλγορίθμου 4.6.

Αλγόριθμος: transition(A, q, a, thres)

Είσοδος: A : το αυτόματο

q : τρέχουσα κατάσταση

a : σελίδα / σύμβολο στην είσοδο

thres : κατώφλι

αρχή

εάν υπάρχει μετάβαση για το a

πήγαινε στην κατάσταση όπου καταλήγει η μετάβαση

αλλιώς

βρες από τις καταστάσεις-παιδιά της q την πιο όμοια ως προς το περιεχόμενο σε σχέση με τη σελίδα a·

εάν η τιμή αυτή > thres

πήγαινε στην κατάσταση αυτή

αλλιώς

πήγαινε στην αρχική κατάσταση του γράφου

τέλος

Αλγόριθμος 4.6 Διαδικασία μετάβασης για τις μεθόδους CANUMGI-A και CANUMGI-B

Ο αλγόριθμος αυτός κάνει την υπόθεση ότι, αν δεν υπάρχει ρητή μετάβαση για κάποια σελίδα a από την τρέχουσα κατάσταση, τότε η κατάσταση-παιδί της που είναι πιο κοντά στη σελίδα a

ως προς το περιεχόμενο αποτελεί την καλύτερη προσέγγιση. Επιπλέον θέτει ένα κατώφλι στην τιμή του μετρικού για να αποφεύγεται η μετάβαση σε κατάσταση τελείως ανόμοια με τη σελίδα a. Στην περίπτωση αυτή επανερχόμαστε στην αρχική κατάσταση. Δηλαδή αγνοούμε τις σελίδες της συνόδου που έχουν εξεταστεί μέχρι στιγμής.

Επομένως, αν υποθέσουμε ότι έχουμε μία σύνοδο χρήσης αποτελούμενη από n σελίδες, $p_1 p_2 ... p_n$, τότε μπορούμε να προσδιορίσουμε την κατάσταση στην οποία αντιστοιχεί η πλοήγηση στις σελίδες αυτές εφαρμόζοντας την παραπάνω μέθοδο διαδοχικά για τις σελίδες από 1 ως n ξεκινώντας από τον αρχικό κόμβο του γράφου.

4.3.1.2 Επιλογή Σελίδων

Το επόμενο βήμα είναι η επιλογή των σελίδων. Θεωρούμε ότι κατάλληλες προς πρόταση σελίδες είναι αυτές που περιέχονται στους κόμβους για τους οποίους υπάρχουν μεταβάσεις από τον τρέχοντα κόμβο. Ωστόσο, κάθε κόμβος είναι μια ομάδα (cluster) από εν γένει πολλές σελίδες. Εδώ τίθενται επομένως δύο ζητήματα: ποιους από τους κόμβους-παιδιά θα χρησιμοποιήσουμε και ποιες σελίδες θα επιλέξουμε από κάθε κόμβο. Για την αντιμετώπιση του πρώτου ζητήματος κάνουμε την υπόθεση ότι μία μετάβαση με υψηλή πιθανότητα σημαίνει ότι οι σελίδες που βρίσκονται στον κόμβο όπου αυτή καταλήγει είναι πιο «κατάλληλες» από ό,τι οι σελίδες ενός κόμβου όπου μεταβαίνουμε μη χαμηλή πιθανότητα. Όσον αφορά το δεύτερο ζήτημα, θεωρούμε ότι μία σελίδα που είναι πιο κοντά στο κέντρο βάρους της ομάδας αντιπροσωπεύει καλύτερα την ομάδα αυτή (δηλαδή χαρακτηρίζει το περιεχόμενό της) και συνεπώς είναι περισσότερο κατάλληλη για να προταθεί στο χρήστη. Η σύγκριση αυτή γίνεται με εφαρμογή του μετρικού του συνημιτόνου ανάμεσα στο διάνυσμα της κάθε σελίδας και στο διάνυσμα της ομάδας όπου η σελίδα περιέχεται.

Στην παρούσα εργασία εφαρμόστηκαν δύο τρόποι επιλογής σελίδων, που ακολουθούν την προσέγγιση που παρουσιάστηκε και κατασκευάζουν μία λίστα προτεινόμενων σελίδων προκαθορισμένου μεγέθους:

- Εύρεση του κόμβου-παιδιού με τη μεγαλύτερη πιθανότητα μετάβασης. Επιλογή των σελίδων που είναι πιο κοντά στο κέντρο βάρους της ομάδας με χρήση του μετρικού συνημιτόνου. Αν οι σελίδες του πρώτου κόμβου δεν επαρκούν, συνεχίζουμε στους επόμενους μέχρι να συμπληρωθεί η λίστα.
- Ταξινόμηση όλων των σελίδων w όλων των παιδιών c του τρέχοντος κόμβου με βάση το γινόμενο της πιθανότητας μετάβασης στον κόμβο c επί την τιμή του μετρικού συνημιτόνου ανάμεσα στην σελίδα w και το κέντρο βάρους του c και επιλογή των καλύτερων.

$$transProb(c) \cdot similarity(w, c)$$

Ο πρώτος τρόπος υπονοεί ότι αρκεί ο έλεγχος του κόμβου με τη μεγαλύτερη πιθανότητα μετάβασης για να εντοπιστούν οι καταλληλότερες σελίδες. Αυτό φαίνεται αληθές στην περίπτωση που η πιθανότητα αυτής της μετάβασης είναι σημαντικά μεγαλύτερη από τις πιθανότητες των άλλων μεταβάσεων. Με το δεύτερο τρόπο, η επιλογή των σελίδων είναι περισσότερο εξισορροπημένη. Στόχος είναι ο εντοπισμός σελίδων που να είναι αρκετά κοντά στο κέντρο βάρους της ομάδας τους και ταυτόχρονα η ομάδα αυτή να αντιστοιχεί σε έναν κόμβο στον οποίο η μετάβαση γίνεται με μεγάλη πιθανότητα από τον τρέχοντα κόμβο.

Με τον ένα ή τον άλλο τρόπο κατασκευάζεται τελικά μία λίστα προτεινόμενων σελίδων ορισμένου μήκους (π.χ. 10) σε φθίνουσα αξιολογική σειρά.

4.3.2 Εξατομικευμένη Πλοήγηση με τη Μέθοδο CANUMGI-C

Στην περίπτωση που γίνεται μείωση διαστασιμότητας, η διαδικασία διαφέρει ελαφρώς από αυτή που παρουσιάστηκε στην προηγούμενη ενότητα. Υπενθυμίζεται ότι εδώ οι μεταβάσεις του αυτομάτου χαρακτηρίζονται από ομάδες σελίδων, όπως έχουν προκύψει από την αρχική ομαδοποίηση, και όχι από χωριστές σελίδες. Επομένως, αν υποθέσουμε ότι διαθέτουμε μία ακολουθία σελίδων $p_1 p_2 ... p_n$ και θέλουμε με βάση αυτές να διασχίσουμε το γράφο, πρέπει πρώτα να μεταβούμε από το επίπεδο των σελίδων σε αυτό των ομάδων. Πρέπει δηλαδή να αντιστοιχίσουμε μονοσήμαντα τις σελίδες του Παγκόσμιου Ιστού στις ομάδες που διαθέτουμε. Επειδή όμως οι ομάδες έχουν κατασκευαστεί από ένα πεπερασμένο σύνολο σελίδων, είναι εύλογο πολλές από τις νέες σελίδες που ελέγχουμε να μην περιέχονται ήδη σε κάποια ομάδα. Αυτές πρέπει συνεπώς να καταταχθούν (classify) σε κάποια ομάδα με βάση κάποιο ευριστικό κριτήριο. Αυτό μπορεί να είναι η ομοιότητα περιεχομένου ανάμεσα στη σελίδα και στην κάθε ομάδα με βάση το διάνυσμα των λέξεων-κλειδιών. Στον Αλγόριθμο 4.7 παρουσιάζεται σχηματικά ο τρόπος με τον οποίο γίνεται αυτή η κατάταξη.

Αλγόριθμος: classifyPage(p, C)

Είσοδος: p : η προς εξέταση σελίδα

C : η υπάρχουσα ομαδοποίηση

Έξοδος: μία ομάδα

αρχή

εάν υπάρχει ήδη η p σε κάποια ομάδα της C

επιστρέφεται η ομάδα αυτή

αλλιώς

υπολόγισε το μετρικό συνημιτόνου ανάμεσα στην p και όλες τις ομάδες της C·

επιστρέφεται η ομάδα στην οποία αντιστοιχεί η μεγαλύτερη τιμή

τέλος

Αλγόριθμος 4.7 Κατάταξη σελίδας σε ομάδα για τη μέθοδο CANUMGI-C

Δεδομένου του τρόπου κατάταξης των σελίδων σε ομάδες, η μετάβαση από κόμβο σε κόμβο του αυτομάτου γίνεται πλέον με τρόπο παρόμοιο με αυτόν που παρουσιάστηκε για τις δύο πρώτες παραλλαγές της μεθόδου. Αν και με τη μείωση διαστασιμότητας το πλήθος των πιθανών συμβόλων έχει μειωθεί σημαντικά, χρειάζεται ακόμα ένας τρόπος διαχείρισης της περίπτωσης που δεν προβλέπεται μετάβαση για κάποιο σύμβολο. Ο αλγόριθμος μετάβασης παρουσιάζεται σχηματικά παρακάτω (Αλγόριθμος 4.8).

Εφαρμόζοντας τον αλγόριθμο αυτό για μια ακολουθία σελίδων, καταλήγουμε σε κάποια κατάσταση, όπως και προηγουμένως. Με βάση αυτή την κατάσταση επιλέγουμε τις σελίδες που θα προτείνουμε με έναν από τους δύο τρόπους που περιγράφονται στην Υποενότητα 4.3.1. Πρέπει όμως να σημειωθεί ότι πλέον ως σελίδες ενός κόμβου θεωρούνται όλες οι σελίδες που περιέχονται σε όλες τις ομάδες που περιέχονται στον κόμβο. Επίσης, το διάνυσμα που χαρακτηρίζει το κέντρο βάρους ενός κόμβου ισούται με το άθροισμα των διανυσμάτων των ομάδων που αυτός περιέχει.

Αλγόριθμος: transition(A, C, q, p, thres)

Είσοδος: A : το αυτόματο

C : η υπάρχουσα ομαδοποίηση

q : τρέχουσα κατάσταση

p : σελίδα / σύμβολο στην είσοδο

thres : κατώφλι

αρχή

ομάδα c = classifyPage(p, C)

εάν υπάρχει μετάβαση για την ομάδα c

πήγαινε στην κατάσταση όπου καταλήγει η μετάβαση

αλλιώς

βρες από τις καταστάσεις-παιδιά της q αυτήν με τη μεγαλύτερη τιμή μετρικού περιεχομένου ως προς τη σελίδα p·

εάν η τιμή αυτή > thres

πήγαινε στην κατάσταση αυτή

αλλιώς

πήγαινε στην αρχική κατάσταση του γράφου

τέλος

Αλγόριθμος 4.8 Διαδικασία μετάβασης για τη μέθοδο CANUMGI-C

5

Πειραματική Αξιολόγηση

Στο κεφάλαιο αυτό παρουσιάζονται τα πειράματα που έγιναν για την αξιολόγηση των μεθόδων, με έμφαση στην ανάδειξη της επίδρασης των διαφόρων παραμέτρων.

5.1 Περιβάλλον των Πειραμάτων

Για τα πειράματα χρησιμοποιήθηκε ένα σύνολο δεδομένων που περιλαμβάνει δεδομένα χρήσης του Παγκόσμιου Ιστού από αρχεία καταγραφής μιας εταιρείας παροχής υπηρεσιών διαδικτύου (ISP). Πρόκειται για 12932 καταγεγραμμένες επισκέψεις σελίδων του Παγκόσμιου Ιστού που έκαναν πελάτες της εταιρείας ISP μέσα σε μικρό χρονικό διάστημα. Τα δεδομένα αυτά έχουν περάσει από μία φάση προεπεξεργασίας, όπου οι παραπάνω καταγεγραμμένες προσπελάσεις χωρίστηκαν σε 1468 συνόδους χρήσης. Οι προσπελάσεις αυτές έγιναν σε 7214 διαφορετικές σελίδες. Επίσης βρέθηκαν οι λέξεις-κλειδιά που χαρακτηρίζουν κάθε διαφορετική σελίδα. Η πληροφορία αυτή για κάθε σελίδα αναπαρίσταται σε μορφή διανύσματος μεγέθους 5086, όσες είναι συνολικά οι λέξεις-κλειδιά που εμφανίζονται στις υπάρχουσες σελίδες.

Στο πλαίσιο μιας διαδικασίας Μηχανικής Μάθησης, το σύνολο των συνόδων χρήσης χωρίστηκε σε δύο μέρη. Το πρώτο, το δείγμα εκπαίδευσης, περιλαμβάνει 983 συνόδους και χρησιμοποιείται για την εκπαίδευση του μοντέλου, ενώ το δεύτερο, το δείγμα ελέγχου, περιλαμβάνει 485 συνόδους και χρησιμοποιείται για την αξιολόγηση του μοντέλου. Στο

δείγμα εκπαίδευσης περιέχονται 4959 διαφορετικές σελίδες, ενώ στο δείγμα ελέγχου 2667 διαφορετικές σελίδες. Από αυτές μόνο οι 412 είναι κοινές και στα δύο σύνολα.

5.2 Κριτήριο και Διαδικασία Αξιολόγησης

Σε εφαρμογές Ανάκτησης Πληροφοριών (Information Retrieval) η αξιολόγηση της ταξινομημένης λίστας αντικειμένων γίνεται συνήθως με δύο μεγέθη, την ανάκληση (recall) και την ακρίβεια (precision). Το πρώτο εκφράζει το ποσοστό των σχετικών αντικειμένων που ανακτήθηκαν και το δεύτερο το ποσοστό των ανακτημένων αντικειμένων που είναι σχετικές. Ωστόσο, σε μία εφαρμογή όπου δεν υπάρχει επίβλεψη δεν είναι σαφές πόσα αντικείμενα (πόσες σελίδες) είναι σχετικά ως προς αυτό που αναζητείται.

Το ενδιαφέρον εδώ εστιάζεται στο πόσο χρήσιμη είναι μία λίστα προτεινόμενων σελίδων στο χρήστη. Για την αξιολόγηση αυτή χρησιμοποιείται η έννοια της αναμενόμενης χρησιμότητας (expected utility) που παρουσιάζεται στο [BHK98]. Η αναμενόμενη χρησιμότητα μιας λίστας είναι σε γενικές γραμμές η πιθανότητα να δει ο χρήστης την εκάστοτε προτεινόμενη σελίδα πολλαπλασιασμένη επί τη χρησιμότητά της. Δεχόμαστε ότι η πιθανότητα επιλογής κάποιου συνδέσμου είναι φθίνουσα ως προς το μήκος της λίστας. Για τον υπολογισμό της χρησιμότητας μιας σελίδας εκμεταλλευόμαστε το μετρικό περιεχομένου, όπως εξηγείται παρακάτω.

Για τη διαδικασία αξιολόγησης χρησιμοποιούμε τις συνόδους χρήσης του δείγματος ελέγχου. Για κάθε μία από αυτές τις συνόδους κρύβουμε την τελευταία σελίδα, έστω w, και ακολουθούμε τη διαδικασία επιλογής σελίδων που περιγράφηκε στο προηγούμενο κεφάλαιο. Η λίστα των σελίδων $a_0, a_1, ..., a_{n-1}$ που προκύπτει είναι ταξινομημένη με κάποιο αξιολογικό κριτήριο. Θεωρούμε ότι η λίστα προτεινόμενων σελίδων είναι τόσο καλύτερη όσο πιο όμοιες με την κρυμμένη σελίδα είναι αυτές ως προς το περιεχόμενό τους. Στη συνέχεια υπολογίζουμε το μέγεθος:

$$EU_w = \sum_{j=0}^{n-1} \frac{similarity\left(w, a_j\right)}{2^{j/h}}$$

Στον παραπάνω τύπο, n είναι το μήκος της λίστας και h είναι ο χρόνος ημίσειας ζωής. Η παράμετρος αυτή είναι ο αύξων αριθμός της σελίδας που έχει πιθανότητα 50% να διαβαστεί. Με αυτό τον τρόπο επιτυγχάνεται εκθετική μείωση της συμβολής της κάθε σελίδας στο άθροισμα, καθώς διατρέχουμε τη λίστα. Η παράμετρος h τίθεται ίση με 5, αφού αυτή η τιμή χρησιμοποιείται ευρέως σε παρόμοιες εφαρμογές. Η ομοιότητα υπολογίζεται με βάση το μετρικό του συνημιτόνου ανάμεσα στα διανύσματα των δύο σελίδων και παίρνει τιμές στο διάστημα [0,1].

Η συνολική αναμενόμενη χρησιμότητα για όλες τις συνόδους χρήσης υπολογίζεται από τον τύπο:

$$EU = 100 \frac{\sum_w EU_w}{\sum_w EU_w^{\max}}$$

όπου $EU_w^{\max}$ είναι η μέγιστη χρησιμότητα που μπορεί να επιτευχθεί για κάποια δεδομένη προτεινόμενη λίστα και για την αντίστοιχη κρυμμένη σελίδα w. Αυτό ισοδυναμεί με το να έχει η w ομοιότητα 1 με κάθε σελίδα της λίστας:

$$EU_w^{\max} = \sum_{j=0}^{n-1} \frac{1}{2^{j/h}}$$

Με τον τρόπο αυτό γίνεται μία κανονικοποίηση της τιμής της αναμενόμενης χρησιμότητας ως προς το μήκος της λίστας προτεινόμενων σελίδων. Παρά την κανονικοποίηση αυτή, το μετρικό φαίνεται να ευνοεί τις μικρές λίστες. Αυτό το χαρακτηριστικό επιτείνεται λόγω της ανομοιογένειας των δεδομένων χρήσης και λόγω του περιορισμένου αριθμού δεδομένων που είναι διαθέσιμα, με αποτέλεσμα να είναι δύσκολο να βρεθούν πολλές σελίδες που να έχουν σχετικά μεγάλη τιμή ομοιότητας περιεχομένου με την κρυμμένη σελίδα.

5.3 Βάση Σύγκρισης

Για να μπορέσουμε να αξιολογήσουμε την επίδοση των μεθόδων Συμπερασμού Γραμματικών στη μοντελοποίηση της πλοήγησης των χρηστών, κατασκευάζουμε πρώτα και αξιολογούμε ένα απλό μοντέλο που δεν κάνει χρήση Συμπερασμού Γραμματικών. Συγκεκριμένα, κάνουμε αρχικά ομαδοποίηση των 4959 σελίδων που περιέχονται στα δεδομένα εκπαίδευσης σε έναν ορισμένο αριθμό ομάδων. Η επιλογή σελίδων και η αξιολόγηση της λίστας προτεινόμενων σελίδων γίνονται ως εξής. Παίρνουμε τις συνόδους χρήσης του δείγματος ελέγχου και αποκρύπτουμε την τελευταία σελίδα, w. Για κάθε σύνοδο ταξινομούμε τις ομάδες ως προς το πόσο όμοιες είναι κατά μέσο όρο με όλες τις σελίδες της συνόδου (εκτός της w) ως προς το μετρικό του συνημιτόνου. Στη συνέχεια, επιλέγουμε τις n καλύτερες σελίδες από την πρώτη ομάδα, δηλ. αυτές που είναι πιο κοντά στο κέντρο βάρους της. Αν η ομάδα αυτή περιέχει λιγότερες από n σελίδες, τότε συνεχίζουμε με τον ίδιο τρόπο στην επόμενη ομάδα. Στα παρακάτω θεωρούμε $n = 10$. Η αξιολόγηση της λίστας προτεινόμενων σελίδων γίνεται με το μέγεθος της αναμενόμενης χρησιμότητας που παρουσιάστηκε προηγουμένως.

Παρακάτω παρουσιάζονται στον Πίνακα 5.1 οι τιμές αξιολόγησης που υπολογίστηκαν για διάφορες περιπτώσεις ομαδοποίησης.

Πίνακας 5.1 Τιμές αξιολόγησης του μοντέλου βάσης σύγκρισης

# ομάδες	Αναμενόμενη Χρησιμότητα
400	16.66
900	19.16
2000	20.48
2500	21.51
3000	22.66
3437	24.25
4959	23.26

Η τελευταία τιμή που αναγράφεται αντιστοιχεί στην περίπτωση που δεν έχει γίνει καθόλου ομαδοποίηση. Σε αυτή την περίπτωση απλώς ταξινομούνται όλες οι σελίδες με τον τρόπο που περιγράφηκε παραπάνω και επιλέγονται οι 10 πρώτες. Η μεγαλύτερη τιμή εμφανίζεται στην περίπτωση της ομαδοποίησης σε 3437 ομάδες. Αυτή αντιστοιχεί στην περίπτωση όπου δύο σελίδες ομαδοποιούνται, όταν χαρακτηρίζονται από ακριβώς το ίδιο διάνυσμα λέξεων-κλειδιών. Δηλαδή δύο σελίδες της ίδιας ομάδας έχουν ακριβώς την ίδια αναπαράσταση και συνεπώς μπορούν να θεωρηθούν ταυτόσημες στο πλαίσιο της συγκεκριμένης εφαρμογής.

5.4 Προσδιορισμός Παραμέτρων

Κατά τη διαδικασία κατασκευής του μοντέλου (Ενότητα 4.2) εμφανίζονται μία σειρά παραμέτρων που πρέπει να προσδιοριστούν. Αυτές συνοπτικά είναι:

1. Κατώφλι μετρικού χρήσης.
2. Τρόπος συνδυασμού των *p*-τιμών στο κριτήριο χρήσης (ελάχιστη ή μέση τιμή των *p*-τιμών).
3. Κατώφλι μετρικού περιεχομένου.
4. Τρόπος συνδυασμού των κριτηρίων χρήσης και περιεχομένου. Για την CANUMGI-A λογική σύζευξη ή διάζευξη. Για την CANUMGI-B (και CANUMGI-C) ελάχιστη τιμή, μέγιστη τιμή ή σταθμισμένο βάρος των δύο μετρικών.
5. Το πλήθος των ομάδων στην περίπτωση μείωσης διαστασιμότητας (μέθοδος CANUMGI-C).
6. Αποκλεισμός ή μη των αυτομεταβάσεων στην περίπτωση μείωσης διαστασιμότητας.

Επιπλέον, υπάρχουν τρεις παράμετροι σχετικές με τη χρήση του μοντέλου (Ενότητα 4.3):

7. Τρόπος επιλογής σελίδων: καλύτερες σελίδες από τον καλύτερο κόμβο-παιδί ή ταξινόμηση όλων των σελίδων ως προς το γινόμενο πιθανότητας μετάβασης και ομοιότητας ως προς το κέντρο της ομάδας.
8. Κατώφλι στη διαδικασία μετάβασης από κόμβο σε κόμβο.
9. Μήκος λίστας προτεινόμενων σελίδων.

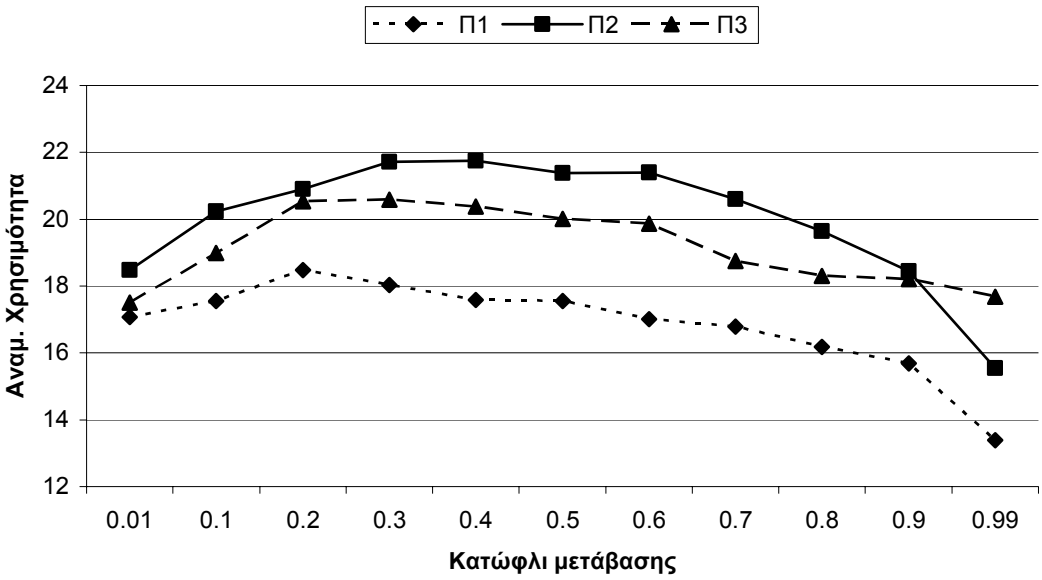
Ο προσδιορισμός των παραμέτρων αρχίζει από αυτές που αναφέρονται στη χρήση του μοντέλου, επειδή επηρεάζουν το μέτρο αξιολόγησης και πρέπει να κρατηθούν σταθερές για να γίνει δυνατή η σύγκριση των διαφόρων μοντέλων. Έπειτα ακολουθεί ο προσδιορισμός των παραμέτρων κατασκευής του μοντέλου.

5.4.1 Προσεγγίσεις στην Επιλογή Σελίδων

Όσον αφορά την επιλογή σελίδων, παρατηρήθηκε ότι η τιμή που επιτυγχάνεται με τον πρώτο τρόπο είναι πάντα 10% - 25% μεγαλύτερη από αυτή με το δεύτερο, ανεξαρτήτως επιλογής μεθόδου και λοιπών παραμέτρων. Αυτό οφείλεται στο γεγονός ότι οι περισσότεροι κόμβοι του επαγόμενου γράφου έχουν μία μετάβαση μεγάλης πιθανότητας (πάνω από 50%) στον εαυτό τους. Δηλαδή ο γράφος τείνει να σχηματίζει ομάδες σελίδων στις οποίες παραμένει κατά μεγάλο ποσοστό ένας χρήστης κατά τη διάρκεια μιας συνόδου. Υπολογίστηκε πράγματι ότι για κατώφλι 0.01 στη διαδικασία μετάβασης (παράμετρος 8) σχεδόν το 50% των συνολικών μεταβάσεων που γίνονται κατά την προσπέλαση των συνόδων χρήσης του δείγματος ελέγχου είναι αυτομεταβάσεις. Επομένως, με την επιλογή σελίδων από τον κόμβο-παιδί με τη μεγαλύτερη πιθανότητα μετάβασης (δηλ. από τον ίδιο τον κόμβο) επιτυγχάνεται μεγαλύτερη χρησιμότητα από ό,τι με την επιλογή από όλους τους κόμβους-παιδιά με έναν ομοιόμορφο τρόπο. Πάντως αυτή η διαδικασία διαφέρει θεωρητικά από την απλή επιλογή σελίδων από ομάδες, που χρησιμοποιείται ως βάση σύγκρισης, καθότι εδώ οι κόμβοι του αυτομάτου έχουν δημιουργηθεί με βάση και τα δεδομένα χρήσης και δεν είναι απλώς ομάδες σελίδων όμοιων ως προς το περιεχόμενο, αν και στην πράξη φαίνεται ότι το κριτήριο ομοιότητας περιεχομένου παίζει τον αποφασιστικότερο ρόλο στη διαδικασία συγχώνευσης των κόμβων. Στη συνέχεια παρουσιάζονται τιμές αναμενόμενης χρησιμότητας που έχουν προκύψει μόνο με τον πρώτο τρόπο επιλογής σελίδων.

5.4.2 Κατώφλι στη Διαδικασία Μετάβασης

Κατά τη διαδικασία μετάβασης από κόμβο σε κόμβο υπάρχει ένα κατώφλι που καθορίζει εάν θα γίνει μετάβαση σε κάποιον κόμβο παιδί ή επάνοδος στον αρχικό κόμβο του γράφου. Στο Σχήμα 5.1 φαίνεται πώς μεταβάλλεται η τιμή της αναμενόμενης χρησιμότητας σε σχέση με το κατώφλι αυτό για διάφορες εκτελέσεις της μεθόδου.



Σχήμα 5.1 Μελέτη κατωφλιού μετάβασης: CANUMGI-B (Π1 και Π2) και CANUMGI-C (Π3)

Οι καμπύλες Π1 και Π2 αφορούν σε γράφους που κατασκευάστηκαν με τη μέθοδο CANUMGI-B αλλά με διαφορετικές τιμές παραμέτρων για το κατώφλι του μετρικού περιεχομένου (0.01 και 0.1 αντίστοιχα) καθώς και για τον τρόπο συνδυασμού των δύο κριτηρίων (βάρος 0.5 και 0.6 αντίστοιχα). Η καμπύλη Π3 αντιστοιχεί σε γράφο κατασκευασμένο με τη μέθοδο CANUMGI-C με αρχική ομαδοποίηση σε 2500 ομάδες (και τιμές 0.1 και 0.5 για τις άλλες δύο παραμέτρους). Επιλέχτηκαν να παρουσιαστούν αυτές οι τρεις περιπτώσεις, διότι αντιστοιχούν σε χαρακτηριστικές τιμές των λοιπών παραμέτρων και έχουν σχετικά καλά αποτελέσματα. Σε όλες τις περιπτώσεις παρατηρείται ότι η αναμενόμενη χρησιμότητα αυξάνεται με την αύξηση του κατωφλιού μετάβασης μέχρι μία τιμή κατωφλιού που κυμαίνεται από 0.2 ως 0.4 και στη συνέχεια φθίνει. Σημειώνεται ότι μεγαλύτερη τιμή του κατωφλιού σημαίνει ότι περισσότερες μεταβάσεις θα γίνονται στον αρχικό κόμβο. Με αυτό τον τρόπο εξασφαλίζεται ότι δε θα γίνονται μεταβάσεις σε κόμβους που είναι τελείως ανόμοιοι ως προς το περιεχόμενο με την τρέχουσα σελίδα της συνόδου. Από την άλλη, αν το κατώφλι τεθεί πολύ υψηλά, τότε η διαδικασία εκφυλίζεται σε συνεχή επάνοδο στον αρχικό κόμβο, με αποτέλεσμα ο γράφος να χάνει τη χρησιμότητά του. Για τον γράφο που αντιστοιχεί στην καμπύλη Π2 υπολογίστηκε ότι για κατώφλι 0.01 το 9% των μεταβάσεων γίνονται στον αρχικό κόμβο, ενώ για κατώφλι 0.3 η τιμή αυτή ανέρχεται στο 47%. Το αποτέλεσμα αυτό υπονοεί ότι για την πρόταση σελίδων αρκεί συνήθως η ύπαρξη πληροφορίας μόνο για τις λίγες προηγούμενες πλοηγήσεις ενός χρήστη και δεν είναι απαραίτητη η γνώση μιας ολόκληρης μακροχρόνιας συνόδου χρήσης. Τέλος, η ομαλότερη πτώση της αναμενόμενης χρησιμότητας στην καμπύλη Π3 οφείλεται στο γεγονός ότι έχει προηγηθεί μείωση διαστασιμότητας. Τα δυνατά σύμβολα είναι πλέον περιορισμένα (2500 στην περίπτωση αυτή), με αποτέλεσμα να υπάρχει μεγαλύτερη πιθανότητα χρήσης μίας ρητής μετάβασης για

κάποιο σύμβολο κατά τη διαδικασία προσπέλασης των σελίδων του δείγματος ελέγχου. Έτσι, μία αύξηση του κατωφλιού μετάβασης δεν αυξάνει σημαντικά το πλήθος των επανόδων στον αρχικό κόμβο του γράφου. Η μέθοδος CANUMGI-C είναι συνεπώς λιγότερο ευαίσθητη στην επιλογή του κατωφλιού μετάβασης.

5.4.3 Μήκος Λίστας Προτεινόμενων Σελίδων

Σε όλα τα αποτελέσματα που παρουσιάστηκαν μέχρι τώρα και σε αυτά που θα ακολουθήσουν θεωρήθηκε ότι η λίστα των σελίδων που προτείνονται στο χρήστη έχει μήκος 10. Όπως αναφέρθηκε στην Ενότητα 5.2, η τιμή του μέτρου αξιολόγησης εξαρτάται από το μήκος της λίστας και συγκεκριμένα ευνοούνται οι μικρότερες λίστες. Κατά συνέπεια, μία μείωση του μήκους της λίστας αναμένεται να επιφέρει αύξηση της αναμενόμενης χρησιμότητας. Όμως με τον ίδιο τρόπο συμπεριφέρεται το μέτρο αξιολόγησης και για το μοντέλο που χρησιμοποιείται ως βάση σύγκρισης, με αποτέλεσμα να αλλάζει η τιμή με την οποία συγκρίνουμε τα αποτελέσματα της μεθόδου CANUMGI. Μελλοντικά θα μπορούσε πάντως να μελετηθεί αναλυτικότερα η επίδραση αυτής της παραμέτρου.

5.4.4 Κατώφλι Μετρικού Χρήσης

Για να ευρεθεί κατάλληλη τιμή για το κατώφλι του μετρικού χρήσης, ερευνήθηκε η κατανομή των τιμών του μετρικού χρήσης που προέκυψαν από τον έλεγχο όλων των ζευγών κόμβων της αρχικής κατάστασης (του PPTA). Η κατανομή αυτή διαφέρει ανάλογα με τον τρόπο συνδυασμού τον *p*-τιμών (παράμετρος 2). Επίσης, επειδή εξαρτάται από τη δομή του αρχικού PPTA, είναι διαφορετικό στην περίπτωση της μείωσης διαστασιμότητας (μέθοδος CANUMGI-C). Παρακάτω παρουσιάζεται η διαδικασία προσδιορισμού της παραμέτρου για την περίπτωση των δύο πρώτων μεθόδων (CANUMGI-A και CANUMGI-B). Στον Πίνακα 5.2 παρουσιάζεται το ποσοστό των ζευγών που έχουν τιμή μετρικού χρήσης σε συγκεκριμένα διαστήματα για την περίπτωση που το μετρικό χρησιμοποιεί την ελάχιστη τιμή των *p*-τιμών και στον Πίνακα 5.3 για την περίπτωση της μέσης τιμής των *p*-τιμών.

Στην πρώτη περίπτωση παρατηρούμε ότι υπάρχει μια μεγάλη αιχμή στο διάστημα (1.2, 1.3] και συγκεκριμένα στην τιμή 1.21306, η οποία αντιστοιχεί στην περίπτωση ελέγχου δύο κόμβων στους οποίους έχει γίνει μετάβαση μία μόνο φορά και η μοναδική μετάβαση που διαθέτουν είναι για διαφορετικά σύμβολα. Το φαινόμενο αυτό παρατηρείται λόγω της μεγάλης ανομοιογένειας των δεδομένων χρήσης, που έχει σαν συνέπεια τα περισσότερα υποδέντρα του PPTA να είναι στην ουσία αλυσίδες κόμβων στους οποίους έχει γίνει μόνο μία φορά μετάβαση. Επειδή αυτές οι περιπτώσεις πρέπει να αποκλειστούν (οι κόμβοι να μη θεωρούνται συμβατοί), το κατώφλι πρέπει να τεθεί πάνω από την τιμή της αιχμής, όχι όμως πολύ υψηλότερα, καθώς το ποσοστό που απομένει είναι πολύ μικρό. Έτσι, αποφασίστηκε το

κατώφλι να τεθεί στην τιμή 1.22. Στη δεύτερη περίπτωση, η μεγάλη αιχμή έχει μετατοπιστεί δεξιότερα, στο διάστημα (1.4, 1.5], και συγκεκριμένα στην τιμή 1.47537. Με το ίδιο σκεπτικό το κατώφλι τέθηκε στην τιμή 1.48.

Πίνακας 5.2 Κατανομή τιμών μετρικού χρήσης (ελάχιστη τιμή *p*-τιμών)

Διάστημα	Ποσοστό
[0, 0.8]	0.80%
(0.8, 1]	0.85%
(1, 1.2]	2.83%
(1.2, 1.3]	94.1%
(1.3, 1.9]	0.16%
(1.9, 2]	1.20%

Πίνακας 5.3 Κατανομή τιμών μετρικού χρήσης (μέση τιμή *p*-τιμών)

Διάστημα	Ποσοστό
[0, 1]	0.1%
(1, 1.2]	0.41%
(1.2, 1.3]	19.9%
(1.3, 1.4]	1.39%
(1.4, 1.5]	75.1%
(1.5, 1.6]	1.04%
(1.6, 1.9]	0.54%
(1.9, 2]	1.51%

Στην περίπτωση της μείωσης διαστασιμότητας (μέθοδος CANUMGI-C), η κατανομή εξαρτάται επιπλέον και από το πλήθος των ομάδων (παράμετρος 5), το οποίο καθορίζει τη δομή του PPTA. Ωστόσο, σε κάθε περίπτωση παρουσιάζει τα ίδια χαρακτηριστικά με τις προηγούμενες κατανομές και ιδιαίτερα τη μεγάλη αιχμή στο ίδιο σημείο. Για παράδειγμα, για την περίπτωση με 2000 ομάδες και συνδυασμό των *p*-τιμών με την ελάχιστη τιμή, η κατανομή φαίνεται στον Πίνακα 5.4. Συνεπώς, και στην περίπτωση της μείωσης διαστασιμότητας τέθηκαν οι ίδιες τιμές κατωφλιού όπως προηγουμένως.

Πίνακας 5.4 Κατανομή τιμών μετρικού χρήσης (ελάχιστη τιμή *p*-τιμών) για 2000 ομάδες

Διάστημα	Ποσοστό
[0, 0.8]	1.52%
(0.8, 1]	1.83%
(1, 1.2]	3.64%
(1.2, 1.3]	90.6%
(1.3, 1.9]	0.27%
(1.9, 2]	2.10%

Σημειώνεται ότι η επιλογή του κατωφλιού με τον τρόπο που παρουσιάστηκε εδώ είναι υποβέλτιστη, αφού η κατανομή αλλάζει, καθώς γίνονται οι συγχωνεύσεις. Γι' αυτό θα ήταν χρήσιμο να αλλάζει δυναμικά η τιμή του κατωφλιού, καθώς προχωρά η επαγωγική διαδικασία.

5.4.5 Τρόπος Συνδυασμού των p*-τιμών στο Κριτήριο Χρήσης*

Οι *p*-τιμές που προκύπτουν από τον έλεγχο ομοιότητας δύο μεταβάσεων καθώς και ως προς την πιθανότητα οι κόμβοι να είναι τερματικοί μπορούν να συνδυαστούν με δύο τρόπους: ελάχιστη τιμή αυτών ή μέση τιμή. Ωστόσο, από τα πειράματα με τις μεθόδους CANUMGI-A και CANUMGI-B προέκυψε ότι ο επαγόμενος γράφος στις περιπτώσεις εφαρμογής της μέσης τιμής των *p*-τιμών για κατώφλι μετρικού χρήσης 1.48 αποτελείται από πολύ λίγους κόμβους (περίπου 10 από 7758 κόμβους που έχει το αρχικό PPTA). Επίσης, για μεγαλύτερες τιμές κατωφλιού, πειράματα με τη μέθοδο CANUMGI-A έδειξαν ότι το μέγεθος του γράφου δεν αυξάνει σημαντικά. Αυτό το φαινόμενο εξηγείται από το γεγονός ότι το μετρικό χρήσης, θεωρώντας τη μέση τιμή των *p*-τιμών, γίνεται ανεκτικό ως προς την ύπαρξη ορισμένων ανόμοιων μεταβάσεων, με αποτέλεσμα να θεωρεί πολύ περισσότερα ζεύγη κόμβων συμβατά. Έτσι, γίνονται πάρα πολλές συγχωνεύσεις και κατά συνέπεια ο τελικός γράφος προκύπτει τόσο γενικευμένος που είναι πλέον ακατάλληλος να μοντελοποιήσει την πλοήγηση των χρηστών. Για το λόγο αυτό, χρησιμοποιείται στη συνέχεια μόνο η ελάχιστη τιμή των *p*-τιμών ως τιμή του μετρικού χρήσης.

5.4.6 Κατώφλι Μετρικού Περιεχομένου

Για τον προσδιορισμό της παραμέτρου αυτής, ερευνήθηκε η κατανομή των τιμών μετρικού περιεχομένου που προκύπτουν από τη σύγκριση όλων των ζευγών κόμβων στον αρχικό γράφο, όπως και στην περίπτωση του μετρικού χρήσης. Στον Πίνακα 5.5 παρουσιάζεται η κατανομή αυτή.

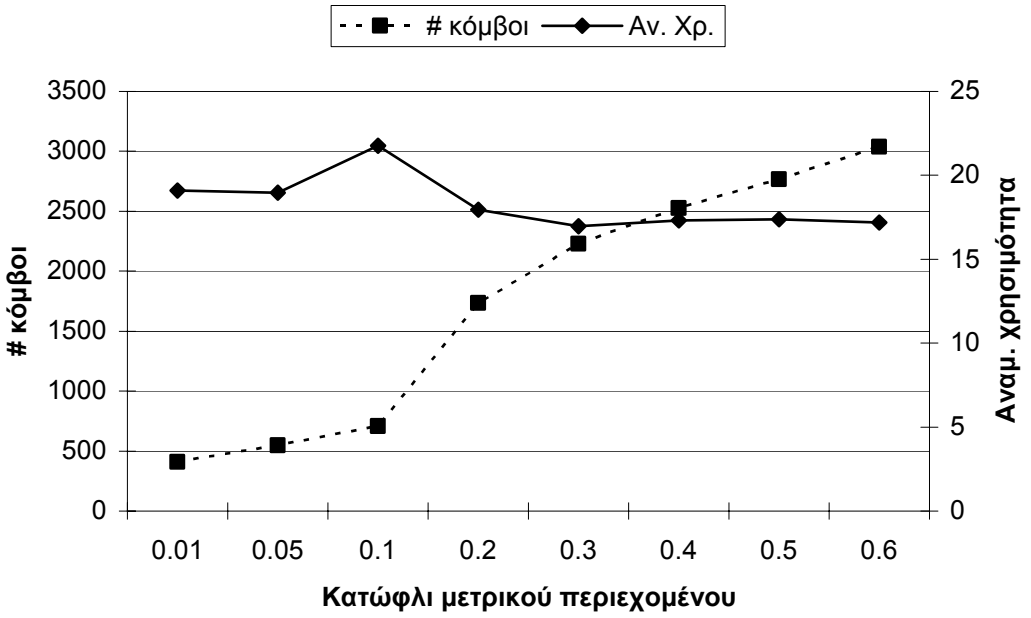
Πίνακας 5.5 Κατανομή τιμών μετρικού περιεχομένου

Διάστημα	Ποσοστό
[0, 0.01]	82.0%
(0.01, 0.05]	9.65%
(0.05, 0.1]	5.70%
(0.1, 0.2]	1.91%
(0.2, 0.3]	0.25%
(0.3, 0.9]	0.20%
(0.9, 1]	0.27%

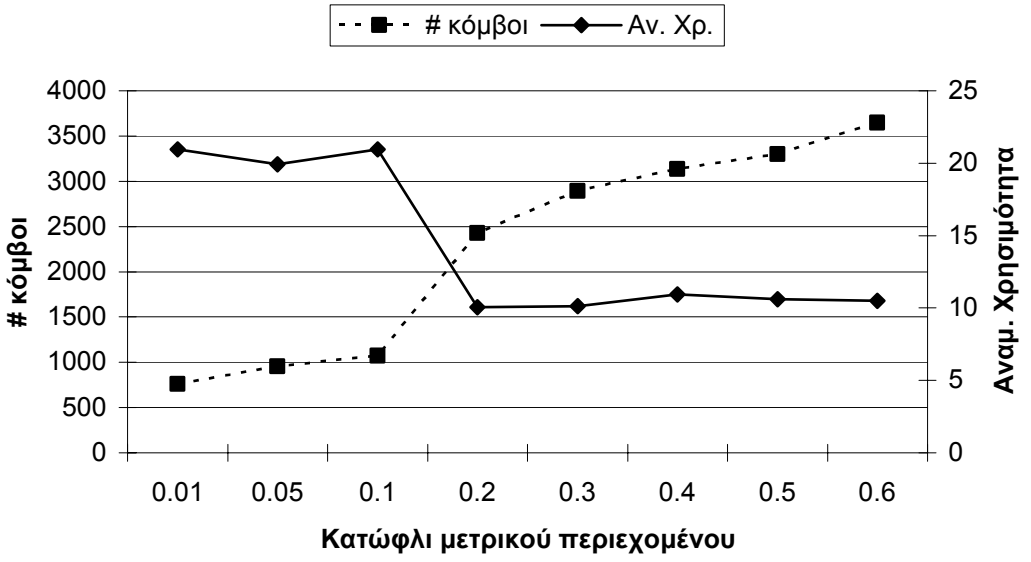
Όπως ήταν αναμενόμενο, λόγω της ανομοιογένειας των δεδομένων χρήσης, τα περισσότερα ζεύγη κόμβων προκύπτουν ανόμοια ως προς το περιεχόμενο και γι' αυτό παρατηρείται η μεγάλη αιχμή στο 0. Ωστόσο, εκτιμάται ότι κατά τη διάρκεια της διαδικασίας κατασκευής του μοντέλου η κατανομή αυτή ομαλοποιείται. Αυτό συμβαίνει, διότι, στην περίπτωση ελέγχου δύο ομάδων σελίδων (αντί για απλές σελίδες), η πιθανότητα αυτές να έχουν τουλάχιστον μία κοινή λέξη-κλειδί είναι αυξημένη, με αποτέλεσμα η τιμή του μετρικού συνημιτόνου να προκύπτει μη μηδενική ακόμα και στην περίπτωση που οι ομάδες αυτές δε θα πρέπει να θεωρηθούν όμοιες. Πάντως, η αρχική κατανομή παραμένει χρήσιμη για τον προσδιορισμό του κατωφλιού. Η τιμή αυτή πρέπει να είναι σίγουρα τουλάχιστον 0.01. Εκτιμάται πάντως ότι πρέπει να είναι ακόμα μεγαλύτερη από την τιμή αυτή, για να αποφευχθεί το ενδεχόμενο δύο κόμβοι να θεωρηθούν εσφαλμένα όμοιοι. Σημειώνεται ότι, όπως και στην περίπτωση του μετρικού χρήσης, μια διαδικασία δυναμικού καθορισμού του κατωφλιού ίσως να ήταν και εδώ χρήσιμη. Στα επόμενα γραφήματα (Σχήματα 5.2 και 5.3) φαίνεται πώς το κατώφλι του μετρικού περιεχομένου επηρεάζει το μέγεθος του τελικού γράφου καθώς και την αναμενόμενη χρησιμότητα.

Και τα δύο γραφήματα αφορούν σε γράφους που έχουν κατασκευαστεί με τη μέθοδο CANUMGI-B. Η διαφορά τους εντοπίζεται στο βάρος του μετρικού περιεχομένου κατά το συνδυασμό των δύο μετρικών (παράμετρος 4), που είναι υψηλότερο στην περίπτωση του Σχήματος 5.2. Και στις δύο περιπτώσεις παρατηρείται ότι το πλήθος των κόμβων αυξάνει με την αύξηση του κατωφλιού, κάτι που είναι αναμενόμενο, αφού το κριτήριο γίνεται έτσι πιο αυστηρό και περιορίζεται το πλήθος των δυνατών συγχωνεύσεων. Ακολούθως, η τιμή της αναμενόμενης χρησιμότητας φαίνεται να μειώνεται μετά από την τιμή κατωφλιού 0.1, διότι τότε τα περιθώρια εύρεσης δύο κόμβων συμβατών ως προς το περιεχόμενο μειώνονται σημαντικά, με συνέπεια ο γράφος να προκύπτει τόσο μεγάλος που δεν είναι πλέον κατάλληλος για μια διαδικασία Ανακάλυψης Προτύπων Διαδοχής. Γενικά, παρατηρήθηκε ότι

καλύτερα αποτελέσματα επιτυγχάνουν οι γράφοι που έχουν μέγεθος περίπου μία τάξη μεγέθους μικρότερο από αυτό του αρχικού PPTA (7758 κόμβοι για τις περιπτώσεις χωρίς μείωση διαστασιμότητας). Η μείωση της τιμής της αναμενόμενης χρησιμότητας είναι



Σχήμα 5.2 Μελέτη κατωφλιού μετρικού περιεχομένου: CANUMGI-B, βάρος μετρ. χρήσης 0.6



Σχήμα 5.3 Μελέτη κατωφλιού μετρικού περιεχομένου: CANUMGI-B, βάρος μετρ. χρήσης 0.8

περισσότερο έντονη στην περίπτωση του Σχήματος 5.3, καθώς το μέγεθος του τελικού γράφου είναι αρκετά μεγαλύτερο από ό,τι στην περίπτωση του Σχήματος 5.2 για την ίδια τιμή κατωφλιού, το οποίο σχετίζεται με τον τρόπο συνδυασμού των δύο μετρικών. Τέλος, φαίνεται

ότι η τιμή κατωφλιού 0.1 είναι γενικά προτιμότερη από την τιμή 0.01, καθώς έτσι αποφεύγεται η συγχώνευση κόμβων που έχουν ελάχιστη ομοιότητα.

Όσον αφορά την περίπτωση που προηγείται μείωση διαστασιμότητας (μέθοδος CANUMGI-C), η αρχική κατανομή των τιμών του μετρικού περιεχομένου εξαρτάται έντονα από την προηγηθείσα ομαδοποίηση των σελίδων. Γενικά, όσο λιγότερες είναι οι ομάδες που δημιουργούνται τόσο περισσότερο ομαλοποιείται η κατανομή, απομακρυνόμενη από την ακραία περίπτωση του Πίνακα 5.5. Για παράδειγμα, στους Πίνακες 5.6 και 5.7 παρουσιάζονται οι κατανομές για μείωση διαστασιμότητας σε 2000 και σε 900 ομάδες αντίστοιχα. Στο πλαίσιο της εργασίας αυτής δεν ήταν εφικτή η αναλυτική μελέτη της επίδρασης του κατωφλιού μετρικού περιεχομένου για κάθε διαφορετική αρχική ομαδοποίηση. Αναζητήθηκε πάντως σε κάθε περίπτωση εκτέλεσης της μεθόδου CANUMGI-C μία τιμή κατωφλιού τέτοια ώστε το ποσοστό των τιμών που την υπερβαίνουν να είναι περίπου ίσο με το αντίστοιχο ποσοστό για κατώφλι 0.1 στην περίπτωση χωρίς μείωση διαστασιμότητας.

Πίνακας 5.6 Κατανομή τιμών μετρικού περιεχομένου για 2000 ομάδες

Διάστημα	Ποσοστό
[0, 0.01]	51.7%
(0.01, 0.05]	21.8%
(0.05, 0.1]	14.1%
(0.1, 0.2]	8.86%
(0.2, 0.3]	2.06%
(0.3, 0.9]	0.24%
(0.9, 1]	1.19%

Πίνακας 5.7 Κατανομή τιμών μετρικού περιεχομένου για 900 ομάδες

Διάστημα	Ποσοστό
[0, 0.01]	29.8%
(0.01, 0.05]	21.3%
(0.05, 0.1]	16.2%
(0.1, 0.2]	17.4%
(0.2, 0.3]	6.28%
(0.3, 0.4]	3.96%
(0.4, 0.9]	1.96%
(0.9, 1]	2.94%

5.4.7 Συνδυασμός Κριτηρίων Χρήσης και Περιεχομένου

Στην περίπτωση της μεθόδου CANUMGI-A, τα δύο μετρικά μπορούν να συνδυαστούν είτε συζευκτικά είτε διαζευκτικά. Όπως ήταν αναμενόμενο, ένας γράφος που προκύπτει με σύζευξη των μετρικών έχει σημαντικά μεγαλύτερο μέγεθος από έναν με διάζευξη, καθώς λιγότεροι κόμβοι θεωρούνται συμβατοί και συνεπώς συγχωνεύσιμοι. Χαρακτηριστικά, για την περίπτωση μέσης τιμής των *p*-τιμών και κατωφλιού περιεχομένου 0.1, με σύζευξη ο επαγόμενος γράφος έχει 2563 κόμβους, ενώ με διάζευξη 9 κόμβους.

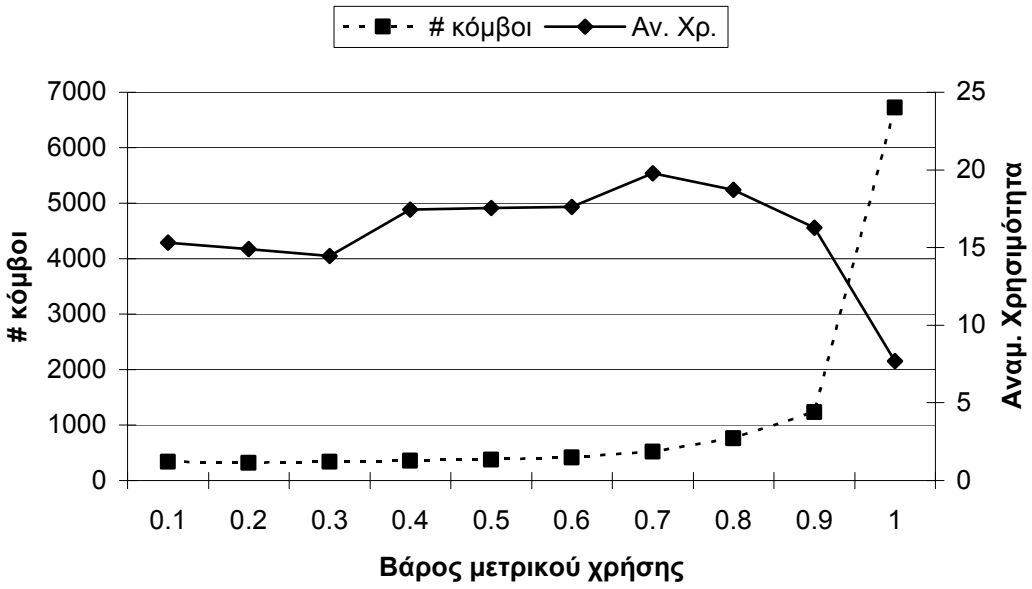
Στην περίπτωση των μεθόδων CANUMGI-B και CANUMGI-C, γίνεται αριθμητική σύγκριση των τιμών που εκφράζουν τα δύο μετρικά. Για την κανονικοποίηση που είναι απαραίτητη έτσι ώστε να γίνει εφικτή η σύγκριση, αξιοποιούνται οι αρχικές κατανομές των τιμών των δύο μετρικών που παρουσιάστηκαν στις Υποενότητες 5.4.4 και 5.4.6. Αρχικά, πρέπει να τεθούν οι τιμές από τα δύο μετρικά σε μία κοινή βάση. Αυτό επιτυγχάνεται με την εξίσωση των κατωφλιών χρήσης και περιεχομένου. Στη συνέχεια, πρέπει να εξομοιωθούν οι διασπορές των κατανομών των τιμών που βρίσκονται πάνω από το κατώφλι του αντίστοιχου μετρικού. Η διαδικασία ξεκινά με τον υπολογισμό ενός παράγοντα κανονικοποίησης:

- Στις δύο κατανομές, απομόνωση των τιμών που είναι πάνω από το αντίστοιχο κατώφλι
- Αφαίρεση από τις τιμές αυτές του αντίστοιχου κατωφλιού
- Υπολογισμός των αντίστοιχων μέσων τιμών (ΜΟ_χρήσης και ΜΟ_περιεχομένου)
- Υπολογισμός του παράγοντα κανονικοποίησης $nf = \text{ΜΟ\_περιεχομένου}/\text{ΜΟ\_χρήσης}$

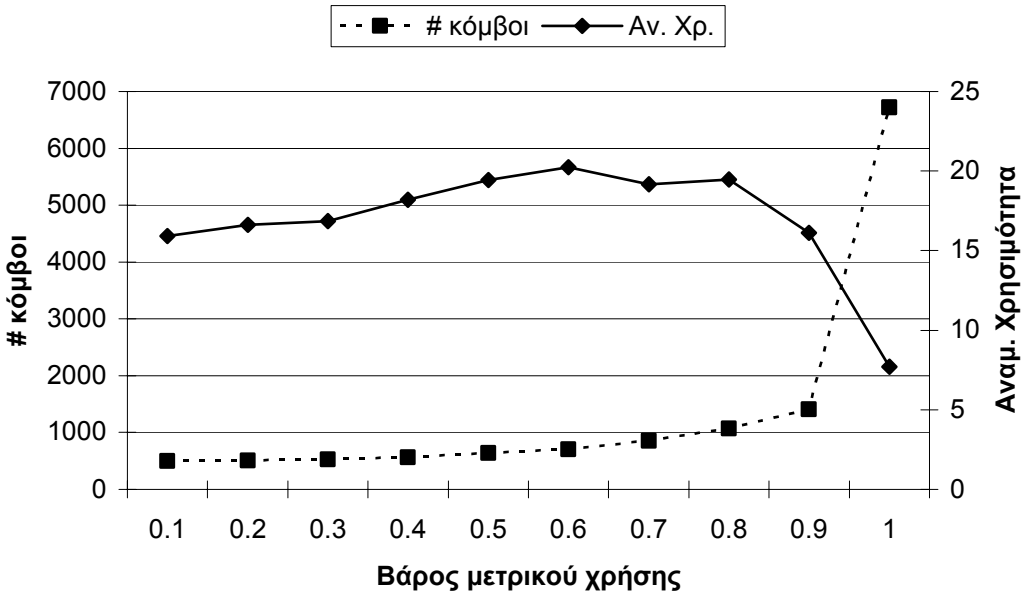
Κατά τον έλεγχο συμβατότητας δύο κόμβων υπολογίζονται οι τιμές μετρικού χρήσης (ΤΜΧ) και μετρικού περιεχομένου (ΤΜΠ). Τελικά συγκρίνονται οι δύο παρακάτω τιμές:

- ΤΜΠ - κατώφλι_περιεχ.
- $nf \cdot (\text{ΤΜΧ - κατώφλι\_χρήσης})$

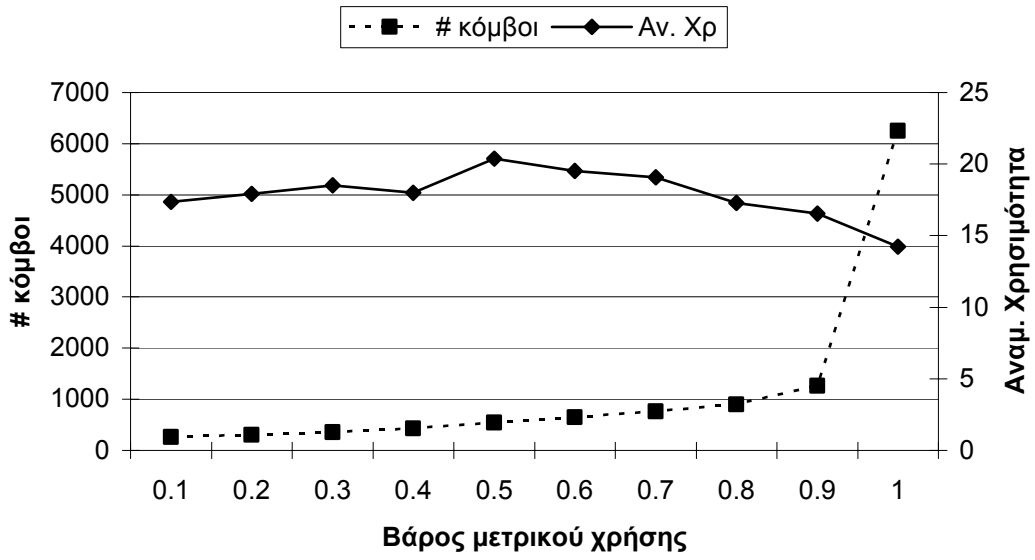
Επειδή ο συνδυασμός των δύο μετρικών με τις συναρτήσεις ελάχιστης και μέγιστης τιμής έχει τα ίδια αποτελέσματα με τις περιπτώσεις σύζευξης και διάζευξης της μεθόδου CANUMGI-A, στη συνέχεια θα ασχοληθούμε μόνο με το σταθμισμένο άθροισμα των δύο μετρικών. Τα παρακάτω σχήματα (Σχήματα 5.4 ως 5.6) δείχνουν πώς επηρεάζονται τα αποτελέσματα από την αύξηση του βάρους που προσδίδεται στο μετρικό χρήσης.



Σχήμα 5.4 Μελέτη συνδυασμού κριτηρίων: CANUMGI-B, κατώφλι μετρικού περιεχ. 0.01



Σχήμα 5.5 Μελέτη συνδυασμού κριτηρίων: CANUMGI-B, κατώφλι μετρικού περιεχ. 0.1



Σχήμα 5.6 Μελέτη συνδυασμού κριτηρίων: CANUMGI-C, 2500 ομάδες, κατώφλι μετρ. περιεχ. 0.1

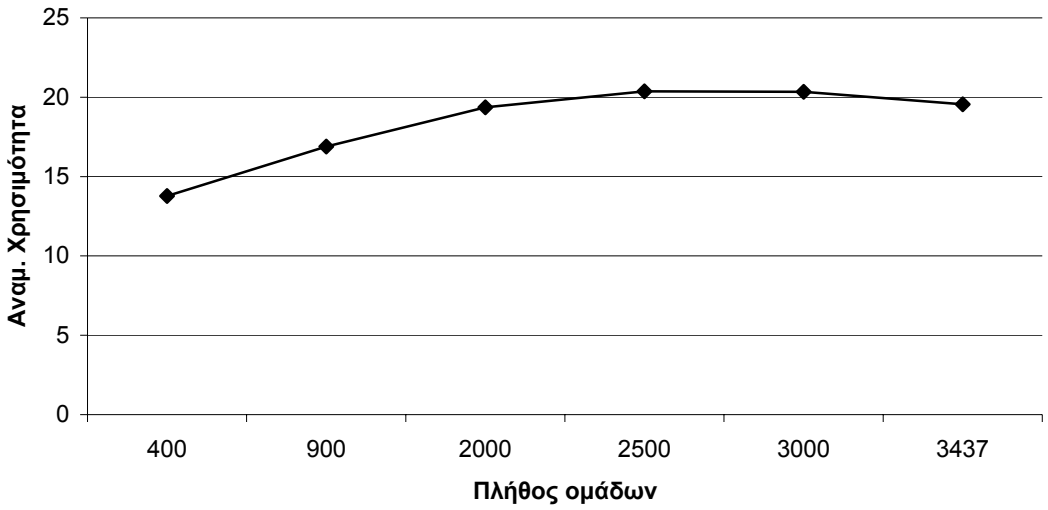
Τα δύο πρώτα γραφήματα αφορούν σε γράφους που έχουν κατασκευαστεί με τη μέθοδο CANUMGI-B. Το πρώτο αντιστοιχεί σε κατώφλι μετρικού περιεχομένου 0.01 και το δεύτερο 0.1. Το τρίτο γράφημα αντιστοιχεί σε γράφους που προέκυψαν από την εκτέλεση της μεθόδου CANUMGI-C με αρχική ομαδοποίηση των σελίδων σε 2500 ομάδες. Σε όλες τις περιπτώσεις παρατηρούμε ότι το μέγεθος του γράφου αυξάνει με την αύξηση του βάρους του μετρικού χρήσης. Αυτό σημαίνει ότι όσο περισσότερο αυξάνεται η συμβολή του μετρικού περιεχομένου τόσο περισσότερες συγχωνεύσεις γίνονται δυνατές, το οποίο οφείλεται στο γεγονός ότι για τα δεδομένα κατώφλια μετρικού περιεχομένου περισσότερα ζεύγη βρίσκονται συμβατά ως προς το μετρικό περιεχομένου παρά ως προς το μετρικό χρήσης, σύμφωνα με τις κατανομές που παρουσιάστηκαν στις Υποενότητες 5.4.4 και 5.4.6. Ως προς την αναμενόμενη χρησιμότητα, από τα δύο πρώτα γραφήματα προκύπτει ότι η μεγαλύτερη τιμή εμφανίζεται για βάρος του μετρικού χρήσης από 0.6 ως 0.8. Το μέγεθος του γράφου στις περιπτώσεις αυτές είναι περίπου μία τάξη μεγέθους μικρότερο από αυτό του αρχικού PPTA. Το γεγονός ότι στο τρίτο γράφημα το μέγιστο εμφανίζεται για λίγο μικρότερη τιμή κατωφλιού οφείλεται στο γεγονός ότι έχει προηγηθεί μείωση διαστασιμότητας και συνεπώς οι κατανομές των τιμών των δύο μετρικών δεν είναι οι ίδιες όπως προηγουμένως.

5.4.8 Πλήθος Ομάδων στη Μείωση Διαστασιμότητας

Στην περίπτωση που προηγείται μείωση διαστασιμότητας (μέθοδος CANUMGI-C), το πλήθος των ομάδων που αρχικά δημιουργούνται επηρεάζει σημαντικά την επίδοση του τελικού μοντέλου. Φάνηκε ότι μία ομαδοποίηση σε λίγες ομάδες μάλλον περιορίζει τη διακριτική ευχέρεια του μοντέλου παρά αναδεικνύει λανθάνουσες ομοιότητες που θα ενίσχυαν τη διαδικασία Ανακάλυψης Προτύπων Διαδοχής. Αυτό επιτείνεται από το γεγονός

ότι τα δεδομένα χρήσης είναι τέτοια ώστε να δημιουργείται συνήθως μία μεγάλη ομάδα από πολλές σελίδες και αρκετές ομάδες με μία μόνο σελίδα (μονοσύνολα).

Η επίδραση του πλήθους των ομάδων είναι δύσκολο να προσδιοριστεί επακριβώς, διότι, όπως ήδη αναφέρθηκε, ο προσδιορισμός του κατωφλιού του μετρικού περιεχομένου εξαρτάται έντονα από την παράμετρο αυτή. Στο Σχήμα 5.7 φαίνεται το πώς μεταβάλλεται η τιμή της αναμενόμενης χρησιμότητας σε σχέση με το πλήθος των ομάδων. Όπως ήδη αναφέρθηκε, ο προσδιορισμός του κατωφλιού μετρικού περιεχομένου εξαρτάται από το πλήθος των ομάδων. Για κάθε περίπτωση ομαδοποίησης έχει επιλεγεί λοιπόν κατάλληλο κατώφλι, έτσι ώστε στην αρχική κατάσταση περίπου το ίδιο πλήθος ζευγών να θεωρούνται συμβατά ως προς το περιεχόμενο. Αυτή η πρόνοια λαμβάνεται, για να γίνουν συγκρίσιμα μεταξύ τους τα αποτελέσματα που προκύπτουν για διαφορετικές αρχικές ομαδοποιήσεις.



Σχήμα 5.7 Επίδραση του πλήθους των ομάδων στην περίπτωση της μεθόδου CANUMGI-C

Το συμπέρασμα που προκύπτει από το γράφημα αυτό είναι ότι μεγάλες τιμές της παραμέτρου (γύρω στο 2500) αντιστοιχούν σε καλύτερες τιμές αξιολόγησης. Ωστόσο, δεν επιτεύχθηκε σε καμία περίπτωση τιμή καλύτερη από τη μέγιστη τιμή που βρέθηκε χωρίς μείωση διαστασιμότητας.

5.4.9 Αποκλεισμός των Αυτομεταβάσεων

Τέλος, αναφέρεται ότι δοκιμάστηκε και η περίπτωση μείωσης διαστασιμότητας με ταυτόχρονη εξαίρεση των μεταβάσεων που αντιστοιχούν σε μεταβάσεις στην ίδια ομάδα κατά τη διαδικασία κατασκευής του αρχικού PPTA. Ωστόσο, τα αποτελέσματα που προέκυψαν αποδείχτηκαν πολύ χειρότερα. Όπως ήδη αναφέρθηκε, στην αρχική περίπτωση ο γράφος τείνει να δημιουργεί ομάδες σελίδων, στις οποίες παραμένει κατά μεγάλο ποσοστό

ένας χρήστης κατά τη διάρκεια μίας συνόδου. Αυτό εκφράζεται με τη μεγάλη πιθανότητα που έχουν οι αυτομεταβάσεις στους κόμβους του γράφου. Με την τελευταία τροποποίηση όμως παρεμποδίζεται η εμφάνιση αυτού του χαρακτηριστικού, με συνέπεια να εξασθενεί το μοντέλο.

5.5 Σύνοψη Παραμέτρων

Από την ανάλυση της προηγούμενης ενότητας προκύπτει ότι ορισμένες παράμετροι μπορούν εύκολα να καθοριστούν εκ των προτέρων. Για παράδειγμα, οι *p*-τιμές στο κριτήριο χρήσης πρέπει να συνδυαστούν με τη συνάρτηση ελάχιστης τιμής. Επίσης, είναι σαφές ότι δεν αποδίδει ο αποκλεισμός των αυτομεταβάσεων στη μείωση διαστασιμότητας. Όσον αφορά την επιλογή των σελίδων, φάνηκε ότι είναι καλύτερα αυτή να γίνεται από τον κόμβο-παιδί όπου καταλήγει η μετάβαση με την υψηλότερη πιθανότητα (και συνήθως συμπίπτει με τον τρέχοντα κόμβο).

Οι υπόλοιπες παράμετροι μπορούν να καθοριστούν μέσα σε κάποια όρια. Σχετικά με το μετρικό χρήσης, αναμένεται να υπάρχει στην κατανομή των αρχικών τιμών μία αιχμή ανεξαρτήτως συνόλου δεδομένων. Επομένως το κατώφλι του μετρικού χρήσης πρέπει να τεθεί υψηλότερα από την αιχμή αυτή. Το ίδιο ισχύει και για το μετρικό περιεχομένου, αν και το εύρος των πιθανών τιμών κατωφλιού που πρέπει να ελεγχθούν είναι στην περίπτωση αυτή μεγαλύτερο. Ο συνδυασμός των δύο μετρικών γίνεται με διάζευξη για την περίπτωση της μεθόδου CANUMGI-A, ενώ για την CANUMGI-B μπορεί να χρησιμοποιηθεί η ελάχιστη τιμή ή ένα σταθμισμένο άθροισμα των δύο μετρικών με βάρος που δε θα παίρνει ακραία τιμή. Όσον αφορά τη μείωση διαστασιμότητας, το μοντέλο αποδίδει καλύτερα με αρχική ομαδοποίηση σε πολλές ομάδες. Τέλος, το κατώφλι στη διαδικασία μετάβασης πρέπει να τεθεί σε μια τιμή στο διάστημα 0.2 ως 0.4.

5.6 Συγκριτική Αποτίμηση Μεθόδων

Η μέθοδος CANUMGI-A δεν έδωσε καλά αποτελέσματα. Η καλύτερη τιμή που επιτεύχθηκε ήταν 8.57 για την περίπτωση που θεωρούμε την χειρότερη τιμή των *p*-τιμών, με κατώφλι μετρικού περιεχομένου 0.05, διάζευξη των δύο μετρικών και κατώφλι στη διαδικασία μετάβασης 0.2. Οι χαμηλές επιδόσεις του CANUMGI-A οφείλονται στη δομή του γράφου που προκύπτει εξαιτίας του τρόπου με τον οποίο ο Alergia επιλέγει τις συγχωνεύσεις. Συγκεκριμένα, δημιουργείται συνήθως ένας μεγάλος κόμβος με πολλές σελίδες και πολλοί κόμβοι με μόνο μία σελίδα.

Τα αποτελέσματα της μεθόδου CANUMGI-B ήταν περισσότερο ενθαρρυντικά. Η δομή του τελικού γράφου είναι ομαλότερη και η επίδοσή του είναι σε κάθε περίπτωση καλύτερη από

ό,τι με την CANUMGI-A. Οι καλύτερες τιμές προέκυψαν για κατώφλι μετρικού περιεχομένου 0.1, βάρος μετρικού χρήσης γύρω στο 0.6 και κατώφλι στη διαδικασία μετάβασης γύρω στο 0.3. Η μεγαλύτερη τιμή αναμενόμενης χρησιμότητας που επιτεύχθηκε ήταν συγκεκριμένα 21.72.

Η μέθοδος CANUMGI-C, που χρησιμοποιεί την ιδέα της μείωσης διαστασιμότητας, δεν έδωσε καλύτερα αποτελέσματα. Ιδιαίτερα για αρχική ομαδοποίηση των σελίδων σε λίγες ομάδες, η αναμενόμενη χρησιμότητα είναι αρκετά χαμηλή. Για μεγαλύτερο πλήθος αρχικών ομάδων τα αποτελέσματα ήταν πάντως συγκρίσιμα με αυτά της μεθόδου CANUMGI-B. Η μεγαλύτερη τιμή επιτεύχθηκε για κατώφλι μετρικού περιεχομένου 0.1, βάρος μετρικού χρήσης 0.5, κατώφλι στη διαδικασία μετάβασης 0.3 και 2500 αρχικές ομάδες και ήταν 20.59.

Πάντως, καμία από τις τρεις μεθόδους δεν επέτυχε αναμενόμενη χρησιμότητα μεγαλύτερη από αυτή που προκύπτει από το απλό μοντέλο που βασίζεται μόνο στην ομοιότητα περιεχομένου και χρησιμοποιείται ως βάση σύγκρισης (24.25). Μάλιστα, από τις τιμές της αναμενόμενης χρησιμότητας που επιτεύχθηκαν από το μοντέλο αυτό για διάφορες περιπτώσεις ομαδοποίησης προέκυψε ότι ούτε η απλή ομαδοποίηση συμβάλλει στην πρόταση σελίδων - η επιλογή των καλύτερων σελίδων από το σύνολο των υπαρχόντων σελίδων αποδίδει καλύτερα. Το αποτέλεσμα αυτό δικαιολογεί ως ένα βαθμό τις χαμηλές επιδόσεις της μεθόδου CANUMGI-C.

6

Επίλογος

Στην πρώτη ενότητα συνοψίζεται η εργασία που έγινε και εκτίθενται τα συμπεράσματα που προέκυψαν. Στη δεύτερη ενότητα δίνονται κάποιες μελλοντικές κατευθύνσεις.

6.1 Σύνοψη και Συμπεράσματα

Στην εργασία αυτή παρουσιάστηκε η μέθοδος CANUMGI (Content-Aware Navigational User Modeling with Grammatical Inference), που μοντελοποιεί την πλοήγηση των χρηστών στον Παγκόσμιο Ιστό με στόχο την παροχή εξατομικευμένων υπηρεσιών στον Ιστό και συγκεκριμένα την πρόβλεψη του επόμενου συνδέσμου ενός χρήστη ή εναλλακτικά την πρόταση σε αυτόν πιθανόν ενδιαφερόντων συνδέσμων. Για το σκοπό αυτό, τροποποιήθηκαν οι μέθοδοι Συμπερασμού Γραμματικών Alergia και Blue Fringe εισάγοντας ιδέες από την περιοχή της Ανάκτησης Πληροφοριών. Επίσης, δοκιμάστηκε η μείωση της διαστασιμότητας του προβλήματος πριν την εφαρμογή μιας μεθόδου συμπερασμού. Στη συνέχεια, τα μοντέλα που προέκυψαν από τις τρεις παραλλαγές της CANUMGI αξιολογήθηκαν με βάση το μετρικό της αναμενόμενης χρησιμότητας και συγκρίθηκαν με την επίδοση ενός μοντέλου που δεν κάνει χρήση Συμπερασμού Γραμματικών.

Καμία από τις τρεις παραλλαγές της CANUMGI δεν κατάφερε να ξεπεράσει σε επιδόσεις το μοντέλο που χρησιμοποιείται ως βάση σύγκρισης, το οποίο στηρίζεται μόνο στην ομοιότητα περιεχομένου των σελίδων. Αυτό πιθανώς σημαίνει ότι η γνώση της σειράς με την οποία ένας χρήστης επισκέπτεται ορισμένες σελίδες του Παγκόσμιου Ιστού δε συμβάλλει στη διαδικασία

πρότασης συνδέσμων, σε αντίθεση με ό,τι ισχύει για την περίπτωση μοντελοποίησης της πλοήγησης σε ένα μόνο ιστοχώρο. Η μεγάλη ανομοιογένεια των δεδομένων χρήσης του Παγκόσμιου Ιστού είναι η κύρια αιτία του γεγονότος αυτού. Πράγματι, οι συγχωνεύσεις των κόμβων κατά την επαγωγική διαδικασία γίνονται εκ των πραγμάτων περισσότερο με βάση την ομοιότητα ως προς το περιεχόμενο παρά ως προς τη χρήση, με αποτέλεσμα να δημιουργούνται μεγάλες ομάδες σελίδων όμοιων ως προς το περιεχόμενο. Γενικά, φαίνεται ότι η πλοήγηση ενός χρήστη στον Παγκόσμιο Ιστό περιορίζεται κατά κύριο λόγο σε ένα σύνολο σελίδων της ίδιας θεματικής κατηγορίας, ενώ οι λίγες μεταβάσεις σε άλλες θεματικές κατηγορίες είναι δύσκολο να προβλεφθούν. Το γεγονός αυτό επιβεβαιώνεται πειραματικά από τη δομή του επαγόμενου γράφου, όπου οι αυτομεταβάσεις έχουν αυξημένη πιθανότητα, ενώ οι μεταβάσεις σε διαφορετικούς κόμβους έχουν μικρή πιθανότητα και μάλλον χαρακτηρίζονται από τυχαιότητα.

Όσον αφορά τις τρεις μεθόδους που αναπτύχθηκαν, προέκυψε ότι η CANUMGI-B, βασισμένη στον αλγόριθμο Blue Fringe, αποδίδει καλύτερα από την CANUMGI-A. Αυτό οφείλεται στο γεγονός ότι η μέθοδος Alergia, στην οποία βασίζεται η CANUMGI-A, επιλέγει με σχετικά αυθαίρετο τρόπο τους κόμβους που θα συγχωνεύσει, με αποτέλεσμα να ευνοείται η δημιουργία ενός κόμβου που περιέχει πάρα πολλές σελίδες και έτσι να καταστρέφεται η δομή του γράφου. Αντίθετα, ο Blue-Fringe επιλέγει με πιο έξυπνο τρόπο τους κόμβους που θα συγχωνεύσει. Τέλος, η τεχνική της μείωσης διαστασιμότητας, που χρησιμοποιείται στη μέθοδο CANUMGI-C, δε συνέβαλε περαιτέρω στη διαδικασία πρότασης σελίδων.

Κατά την πειραματική αξιολόγηση ελέγχθηκαν διάφορες παράμετροι σχετικά με τα κατώφλια των μετρικών, τον τρόπο συνδυασμού αυτών, τη διαδικασία επιλογής σελίδων καθώς και με τη μείωση διαστασιμότητας. Προέκυψε ότι πολλές παράμετροι μπορούν να καθοριστούν εκ των προτέρων, ενώ για τις υπόλοιπες το εύρος των πιθανών τιμών μπόρεσε να περιοριστεί.

Από τη μελέτη της επίδρασης του κατωφλιού της διαδικασίας μετάβασης προέκυψε ότι η γνώση των τελευταίων μόνο βημάτων ενός χρήστη είναι περισσότερο χρήσιμη από τη γνώση της πιο μακροχρόνιας συμπεριφοράς του. Αυτό επιβεβαιώνει το βασικό συμπέρασμα, αλλά αφήνει επίσης ανοικτό το ενδεχόμενο μίας καλύτερης μοντελοποίησης που να χρησιμοποιεί πιο επιλεκτικά τα δεδομένα χρήσης σε συνδυασμό με ένα θεματικό μοντέλο βασισμένο στο περιεχόμενο των σελίδων.

6.2 Μελλοντικές Κατευθύνσεις

Από τα αποτελέσματα της εργασίας αυτής προέκυψαν κάποια θέματα που αξίζει να μελετηθούν στο μέλλον. Καταρχάς μπορεί να δοκιμαστεί ο δυναμικός προσδιορισμός των κατωφλιών μετρικού χρήσης και περιεχομένου όσο εξελίσσεται η διαδικασία επαγωγικού

συμπερασμού για τη βελτιστοποίηση των παραμέτρων αυτών. Επίσης, πρέπει να ελεγχθεί η επίδραση του μήκους της λίστας των προτεινόμενων σελίδων στην τιμή της αναμενόμενης χρησιμότητας καθώς και η αξιοπιστία του ίδιου του μέτρου αξιολόγησης. Η επίδοση της μεθόδου CANUMGI πρέπει επιπλέον να ελεγχθεί σε άλλα σύνολα δεδομένων για να επιβεβαιωθούν τα συμπεράσματα που προέκυψαν από την παρούσα εργασία.

Η μοντελοποίηση της πλοήγησης στον Παγκόσμιο Ιστό και με άλλες προσεγγίσεις, ενδεχομένως απλούστερες, φαίνεται να έχει ενδιαφέρον. Συγκεκριμένα, μπορούν να εφαρμοστούν μαρκοβιανά μοντέλα, μαρκοβιανές αλυσίδες καθώς και bigram models. Επίσης, μπορεί να γίνει μοντελοποίηση περιορισμένων φαινομένων με επιλεκτική χρησιμοποίηση των δεδομένων χρήσης, σε συνδυασμό με ένα προεπιλεγμένο μοντέλο που να βασίζεται στο περιεχόμενο. Μπορεί επιπλέον να δοκιμαστεί η μείωση διαστασιμότητας (όπως στην CANUMGI-C) ακολουθούμενη από την κατασκευή ενός πιθανοτικού αυτομάτου υπερκειμένου (hypertext probabilistic automaton), το οποίο μπορεί στη συνέχεια να χρησιμοποιηθεί για την πρόταση σελίδων. Τέλος, μπορεί να συνδυαστεί ένα μοντέλο που χρησιμοποιεί μείωση διαστασιμότητας και αποκλεισμό των αυτομεταβάσεων με ένα μοντέλο βασισμένο μόνο στο περιεχόμενο.

7

Βιβλιογραφία

[AZN99] Albrecht, D. W., Zukerman, I. and Nicholson, A. E.: *Pre-sending Documents on the WWW: A Comparative Study*, Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI99), (2), Stockholm, pp. 1274-1279, 1999

[Bes95] Bestavros, A.: *Using Speculation to Reduce Server Load and Service Time on the WWW*, Proceedings of CIKM'95: The 4th ACM International Conference on Information and Knowledge Management, Baltimore, Maryland, pp. 403-410, 1995

[BHK98] Breese, J.S., Heckerman, D. and Kadie, C.: *Empirical Analysis of Predictive Algorithms for Collaborative Filtering*, Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, 1998

[BL99] Borges, J. and Levene, M.: *Data Mining of User Navigation Patterns*, Proceedings of Workshop on Web Usage Analysis and User Profiling (WEBKDD) in conjunction with ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA., pp. 31-36, 1999

[CHM+00] Cadez, I., Heckerman, D., Meek, C., Smyth, P. and White, S.: *Visualization of Navigation Patterns on a Web Site Using Model Based Clustering*, Technical Report MSR-TR-00-18, Microsoft Research, 2000

[CO94] Carrasco, R. and Oncina, J.: *Learning Regular Grammars by Means of a State Merging Method*, Proceedings of the ICGI, 1994

[dlH05] de la Higuera, C.: *A Bibliographical Study of Grammatical Inference*, Pattern Recognition 38, pp. 1332-1348, 2005

[Dup94] Dupont, P.: *Regular grammatical inference from positive and negative samples by genetic search: the gig method*, 1994

[Dup96] Dupont, P.: *Incremental regular inference*, Proceedings of the Third ICGI-96, Lecture Notes in Artificial Intelligence, no. 1147, pp. 222-237, Springer-Verlag, 1996

[Dup97] Dupont, P.: *Grammatical Inference: formal and heuristic methods*, Lecture Notes, 1997

[FB75] Fu, K.S. and Booth, T.L.: *Grammatical inference: Introduction and survey. Part I and II*, IEEE Transactions on Syst. Man. and Cybern, 5, pp. 59-72 and 409-423, 1975

[FB92] Frakes, W.B. and Baeza-Yates, R.: *Information Retrieval: Data Structures and Algorithms*, Prentice Hall, 1992

[GMC+92] Giles, C., Miller, D., Chen, D., Chen, H., Sun, G. and Lee, Y.: *Learning and extracting finite state automata with second-order recurrent neural networks*, Neural Computation, 4(3), pp. 393-405, 1992

[Gol67] Gold, E.M.: *Language identification in the limit*, Information and Control, 10 (5), pp. 447-474, 1967

[GV90] García, P. and Vidal, E.: *Inference of K-testable languages in the strict sense and applications to syntactic pattern recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 12 (9), pp. 920-925, 1990

[HK01] Han, J. and Kamber, M.: *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2001

[Kar03] Καραμπατζιάκης Ν.: *Προσαρμογή και Εφαρμογή Μεθόδων Επαγωγικής Κατασκευής Γραμματικών σε Δεδομένα Χρήσης του Παγκόσμιου Ιστού*, Πτυχιακή εργασία, Τμήμα Πληροφορικής, Ε.Κ.Π.Α, 2003

[KMT97] Koshiba, T., Mäkinen, E. and Takada, Y.: *Learning deterministic even linear languages from positive examples*, Theoretical Computer Science, 185 (1), pp. 63-79, 1997

[LPP98] Lang, K.J., Pearlmutter, B.A. and Price, R.: *Results of the Abbadingo One DFA Learning Competition and a New Evidence Driven State Merging*

Algorithm, Proceeding of the Fourth International Colloquium on Grammatical Inference (ICGI-98), 1998

[Mit97] Mitchell, T.: *Machine Learning*, McGraw-Hill, 1997

[OG92] Oncina, J. and García, P.: *Inferring regular languages in polynomial update time*, In Pérez de la Blanca, N., Sanfeliu, A. and Vidal E., editors, Pattern Recognition and Image Analysis, vol. 1, pp. 49-61, World Scientific, 1992

[OS01] Oliveira de, A.L. and Silva, J.P.M.: *Efficient algorithms for the inference of minimum size DFAs*, Machine Learning Journal 44 (1), pp. 93-119, 2001

[PH00] Parekh, R. and Honavar, V.: *Grammar Inference, Automata Induction and Language Acquisition*, Invited chapter in Handbook of Natural Language Processing, Dale, R., Moisl, H. and Somers, H. (editors). Marcel Dekker, New York, 2000

[PP99] Pitkow, J. and Pirolli, P.: *Mining longest repeating subsequences to predict WWW surfing*, Proceedings of the 1999 USENIX User Annual Technical Conference, pp. 139-150, 1999

[PP05] Pierrakos, D. and Paliouras, G.: *Exploiting Probabilistic Latent Information for the Construction of Community Web Directories*, in Ardissono, L., Brna, P. and Mitrovic, A., editors, UM 2005, LNAI 3538, pp. 89-98, Springer-Verlag, 2005

[PPK+00] Paliouras, G., Papatheodorou, C., Karkaletsis, V. and Spyropoulos, C. D.: *Clustering the Users of Large Web Sites into Communities*, Proceedings of International Conference on Machine Learning (ICML), Stanford, California, pp. 719-726, 2000

[PPP+03] Pierrakos, D., Paliouras, G., Papatheodorou, C. and Spyropoulos, C.D.: *Web Usage Mining as a Tool for Personalization: a Survey*, User Modeling and User-Adapted Interaction Journal, vol. 13, issue 4, pp. 311-372, 2003

[RV88] Rulot, H. and Vidal, E.: *An efficient algorithm for the inference of circuit-free automata*, in Ferratè, G., Pavlidis, T., Sanfeliu, A. and Bunke, H., editors, Advances in Structural and Syntactic Pattern Recognition, pp. 173-184, Springer-Verlag, 1988

[Sar00] Sarukkai, R. R.: *Link Prediction and Path Analysis Using Markov Chains*, Proceedings of the 9th World Wide Web Conference, Amsterdam, 2000

[SFW99] Spiliopoulou, M., Faulstich, L. C. and Wilkler, K.: *A data miner analyzing the navigational behavior of Web users*, Proceedings of the Workshop on Machine Learning in User Modeling of the ACAI99, Chania, Greece, pp.

54-64, 1999

[SK99] Sakakibara, Y. and Kondo, M.: *Ga-based learning of context-free grammars using tabular representations*, Proceedings of 16th International Conference on Machine Learning (ICML99), pp. 354-360, 1999

[Tak88] Takada, Y.: *Grammatical Inference for even linear languages based on control sets*, Information Processing Letters, 28 (4), pp. 193-199, 1988

[TDH00] Thollard, F., Dupont, P. and de la Higuera, C.: *Probabilistic DFA inference using Kullback-Leibler divergence and minimality*, Proceedings of the 17th International Conference on Machine Learning, pp. 975-982, 2000

[WA02] Wang, Y. and Acero, A.: *Evaluation of spoken language grammar learning in the atis domain*, Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 2002

[Zhu01] Zhu, T.: *Using Markov Chains for Structural Link Prediction in Adaptive Web Sites*, UM 2001, LNAI 2109, pp 298-300, 2001