Ανακάλυψη Προτύπων Πλοήγησης Χρησιμοποιώντας Εκμάθηση Γραμματικών

Νικόλαος Καραμπατζιάκης*

Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών,
Τμήμα Πληροφορικής και Τηλεπικοινωνιών,
Πανεπιστημιούπολη, 157 84 Αθήνα
email: N.Karampatziakis@di.uoa.gr

Περίληψη Παρουσιάζουμε μια μέθοδο για τη μοντελοποίηση της πλοήγησης των χρηστών ενός δικτυακού τόπου χρησιμοποιώντας εκμάθηση στοχαστικών κανονικών γραμματικών. Με αυτή τη μέθοδο επιτυγχάνουμε καλύτερα μοντέλα από τις Μαρκοβιανές αλυσίδες πρώτης τάξης, οι οποίες έχουν χρησιμοποιηθεί προηγουμένως για τον ίδιο σκοπό, όσον αφορά την ακρίβεια των προβλέψεων και την χρησιμότητα των προτάσεων προς το χρήστη. Για να πετύχουμε συγκρίσιμα αποτελέσματα, εφαρμόζουμε τους ίδιους αλγορίθμους εκμάθησης γραμματικών σε Μαρκοβιανές αλυσίδες, μοντελοποιημένες ως πιθανοτικά αυτόματα. Τα αυτόματα που προκύπτουν κατ΄ αυτό τον τρόπο έχουν καλύτερες επιδόσεις από τις αρχικές Μαρκοβιανές αλυσίδες, ως μοντέλα πλοήγησης των χρηστών, αλλά είναι σημαντικά κατώτερα από τα αυτόματα που προκύπτουν από τις παραδοσιακές μεθόδους εκμάθησης γραμματικών.

1 Εισαγωγή

Σε αυτή τη δουλειά ασχολούμαστε με τη μοντελοποίηση της συμπεριφοράς των χρηστών στον Ιστό. Μας ενδιαφέρει ειδικότερα η ανακάλυψη προτύπων πλοήγησης με τη βοήθεια μεθόδων εκμάθησης γραμματικών (grammatical inference). Όταν ένας χρήστης ζητά μια σελίδα από ένα συγκεκριμένο δικτυακό τόπο, αυτή η δοσοληψία καταγράφεται σε διάφορα σημεία όπως το ιστορικό του browser, το αρχείο απογραφής (log file) του εξυπηρέτη και σε άλλους ενδιάμεσους όπως οι proxies. Αυτές οι δοσοληψίες καλούνται γενικά δεδομένα χρήσης του Παγκοσμίου Ιστού (ΔΧΠΙ). Η εξόρυξη από ΔΧΠΙ είναι ένα σχετικά νέο ερευνητικό πεδίο εμπνευσμένο από την πρόσφατη ανάπτυξη του Παγκοσμίου Ιστού, το οποίο προσπαθεί να αντιμετωπίσει τόσο παραδοσιακά προβλήματα της εξόρυξης δεδομένων όσο και έμφυτες δυσκολίες στην συγκεκριμένη περιοχή, όπως η αξιοπιστία των δεδομένων χρήσης. Δύο διεξοδικές επισκοπήσεις στο πεδίο της εξόρυξης από ΔΧΠΙ μπορούν να βρεθούν στα [16,19]. Στην δική μας περίπτωση εξετάζουμε ΔΧΠΙ στη μορφή ενός αρχείου απογραφής ενός εξυπηρέτη. Από αυτά τα δεδομένα προσπαθούμε να

* Επιβλέποντες: Παναγιώτης Σταματόπουλος (Τμήμα Πληροφορικής και Τηλεπικοινωνιών) και Γεώργιος Παλιούρας (Ινστιτούτο Πληροφορικής και Τηλεπικοινωνιών ΕΚΕΦΕ «Δημόκριτος»)

ανακαλύψουμε ενδιαφέροντα πρότυπα υποθέτοντας ότι η πλοήγηση ενός χρήστη διέπεται από τους κανόνες μιας άγνωστης τυπικής γλώσσας. Για να συμπεράνουμε τη γραμματική της γλώσσας, κάθε ακολουθία ιστοσελίδων που αιτείται ένας χρήστης κατά τη διάρκεια μιας αλληλεπίδρασης με το δικτυακό τόπο, θεωρείται ως ένα θετικό παράδειγμα μιας συμβολοσειράς που ανήκει στην άγνωστη γλώσσα.

Το κίνητρο για την επιλογή μιας γραμματικής για τη μοντελοποίηση της συμπεριφοράς πλοήγησης των χρηστών βρίσκεται στην ακολουθιακή φύση του browsing. Οι χρήστες συνήθως ξεκινούν από κάποια σελίδα, περνούν λίγο χρόνο διαβάζοντάς τη και στη συνέχεια επιλέγουν το σύνδεσμο με τον οποίο πιστεύουν ότι θα ικανοποιήσουν τις πληροφοριακές τους ανάγκες. Η διαδικασία τότε επαναλαμβάνεται, παράγοντας μια ακολουθία από σελίδες που έχει ζητήσει ο κάθε χρήστης. Επομένως, η χρήση τυπικών γλωσσών φάνηκε λογική καθώς είναι κατάλληλες για τη μοντελοποίηση ακολουθιών και επίσης επειδή έχουν ισχυρή μαθηματική θεμελίωση. Για παράδειγμα, οι τυπικές γλώσσες έχουν ήδη χρησιμοποιηθεί για τη μοντελοποίηση βιολογικών ακολουθιών με πολύ καλά αποτελέσματα [13]. Σύμφωνα με όσα γνωρίζουμε, οι υπάρχουσες τεχνικές προς την κατεύθυνσή μας δεν έχουν ως τώρα χρησιμοποιήσει μεθόδους εκμάθησης γραμματικών για την κατασκευή του μοντέλου πλοήγησης των χρηστών. Δημοφιλείς προσεγγίσεις περιλαμβάνουν Μαρκοβιανά μοντέλα, συνήθως πρώτης τάξης [2,3,7,17], ή κατά περίπτωση (ad-hoc) μεθόδους [14,18]. Ωστόσο, οι Μαρκοβιανές αλυσίδες παράγουν απλουστευμένα μοντέλα για την συμπεριφορά πλοήγησης των χρηστών, συναθροίζοντας απλώς την πληροφορία στα δεδομένα, χωρίς να εμπλέκουν καμία διαδικασία εκμάθησης. Από την άλλη, οι κατά περίπτωση προσεγγίσεις συνήθως δεν έχουν κανένα θεωρητικό υπόβαθρο.

Το υπόλοιπο άρθρο οργανώνεται ως εξής. Στην ενότητα 2 αναλύονται οι μέθοδοι εκμάθησης που χρησιμοποιήθηκαν και οι προσαρμογές που έγιναν για το συγκεκριμένο πρόβλημα. Στην ενότητα 3, περιγράφεται το πλαίσιο μέσα στο οποίο διεξήχθησαν τα πειράματα και ερμηνεύονται τα αποτελέσματα τους. Τέλος, στην ενότητα 4 συνάγονται τα συμπεράσματα και προτείνονται μερικές κατευθύνσεις για μελλοντική έρευνα.

2 Εκμάθηση γραμματικών για εξόρυξη ακολουθιών

Σε αυτή τη δουλειά η πλοήγηση του χρήστη σε ένα δικτυακό τόπο μοντελοποιείται σαν να διέπεται από μια άγνωστη στοχαστική κανονική γλώσσα L. Κάθε αλληλεπίδραση του χρήστη με το δικτυακό τόπο μπορεί ισοδύναμα να εκφραστεί ως μια ακολουθία από ιστοσελίδες που ζήτησε ο χρήστης. Αυτές οι ακολουθίες αναπαρίστανται ως συμβολοσειρές που ανήκουν στην L. Έτσι, κάθε ιστοσελίδα που εμφανίζεται στο αρχείο απογραφής του εξυπηρέτη αναπαρίσταται από ένα μοναδικό τερματικό σύμβολο. Επομένως, ένας επισκέπτης του δικτυακού τόπου παρέχει έμμεσα μια συμβολοσειρά που ανήκει στην L ακολουθώντας τους συνδέσμους στις ιστοσελίδες μέχρι να βρει την επιθυμητή πληροφορία ή μέχρι να φύγει.

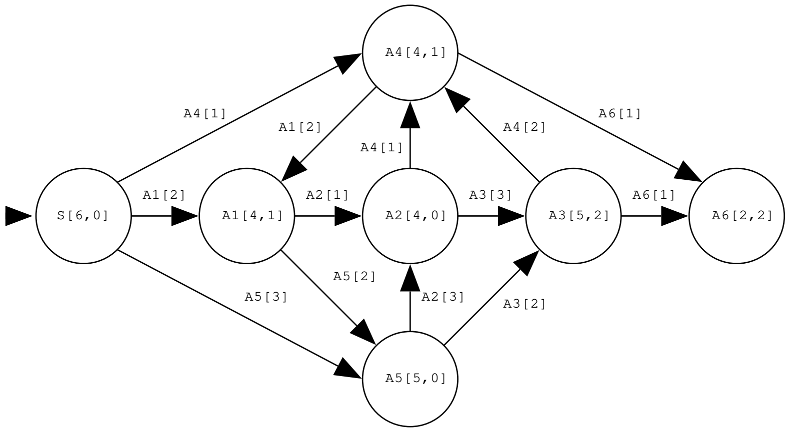
Για την εκμάθηση της γλώσσας L, από τα δεδομένα χρήσης, εφαρμόσαμε τον ευρέως χρησιμοποιούμενο αλγόριθμο Alergia, χρησιμοποιώντας το στατιστικό έλεγχο Hoeffding [8], και μια μικρή τροποποίησή του, χρησιμοποιώντας τον έλεγχο αναλογιών [10]. Η συμβολή μας έγκειται κυρίως στην αρχική υπόθεση με την οποία αυτοί οι αλγόριθμοι ξεκινούν. Δοκιμάστηκαν δύο αρχικές υποθέσεις: το καθιερωμένο πιθανοτικό δενδρικό αυτόματο προθεμάτων (probabilistic prefix tree automaton, PPTA) [15] και το λεγόμενο πιθανοτικό αυτόματο υπερκειμένου (hypertext probabilistic automaton, HPA) [5]. Το τελευταίο έχει χρησιμοποιηθεί παλαιότερα στην εξόρυξη από ΔΧΠΙ για την εύρεση των συχνότερων πορειών σε ένα δικτυακό τόπο [4] αλλά όχι μέσα στο πλαίσιο της εκμάθησης γραμματικών.

Πίνακας 1. Δείγμα συνόλου από συνόδους χρηστών.

Σύνοδος	Ακολουθία σελίδων
1	$A_1 \rightarrow A_2 \rightarrow A_3 \rightarrow A_4$
2	$A_1 \rightarrow A_5 \rightarrow A_3 \rightarrow A_4 \rightarrow A_1$
3	$A_5 \rightarrow A_2 \rightarrow A_4 \rightarrow A_6$
4	$A_5 \rightarrow A_2 \rightarrow A_3$
5	$A_5 \rightarrow A_2 \rightarrow A_3 \rightarrow A_6$
6	$A_4 \rightarrow A_1 \rightarrow A_5 \rightarrow A_3$

Το HPA είχε προταθεί ως ένας συνοπτικός τρόπος μοντελοποίησης των αλληλεπιδράσεων των χρηστών με το δικτυακό τόπο. Αυτή η αλληλεπίδραση αρχικά καταγράφεται στο αρχείο απογραφής του εξυπηρέτη. Στη συνέχεια, η ακολουθία των σελίδων που αιτείται κάθε χρήστης μέσα σε συγκεκριμένο χρονικό διάστημα, η οποία καλείται μια σύνοδος του χρήστη, εξάγεται από αυτό το αρχείο (βλ. πίνακα 1). Δεδομένου ενός συνόλου από συνόδους χρηστών, ένα HPA κατασκευάζεται ως εξής: Για κάθε σελίδα A_i, που εμφανίζεται στις συνόδους, δημιουργείται μια μοναδική κατάσταση στο HPA. Αν κάποιος χρήστης ζητήσει τη σελίδα A_j αμέσως μετά από τη σελίδα A_i, δημιουργείται μια μετάβαση στο HPA από την A_i στην A_j, η οποία μαρκάρεται με το σύμβολο A_j. Επιπλέον, υπάρχει μια αρχική κατάσταση S και μεταβάσεις από την S στις καταστάσεις που αντιστοιχούν σε σελίδες οι οποίες εμφανίζονται πρώτες σε τουλάχιστον μία σύνοδο. Η πιθανότητα κάθε μετάβασης, καθώς και η πιθανότητα κάθε κατάστασης να είναι τελική, υπολογίζεται άμεσα από τις αντίστοιχες συχνότητες στο δεδομένο σύνολο συνόδων. Για παράδειγμα το HPA που μοντελοποιεί τις συνόδους του πίνακα 1, φαίνεται στο σχήμα 1. Από αυτό το σχήμα και την παραπάνω περιγραφή, γίνεται προφανές ότι το HPA είναι ισοδύναμο με Μαρκοβιανές αλυσίδες πρώτης τάξης. Παρακάτω θα αναφερόμαστε εναλλακτικά είτε σε HPA είτε σε Μαρκοβιανές αλυσίδες.

Η προσέγγιση με το HPA ελαττώνει το χώρο αναζήτησης που εξερευνά ο αλγόριθμος εκμάθησης για την εύρεση του αυτόματου στόχου. Επιπλέον, το HPA μπορεί να παραχθεί από το PPTA, συγχωνεύοντας τις καταστάσεις που αντιστοιχούν στις ίδιες σελίδες. Επομένως, το HPA είναι μια γενίκευση του αντίστοιχου PPTA. Αυτό σημαίνει ότι ένα μεγάλο μέρος του αρχικού χώρου αναζήτησης, το οποίο περιέχει πιο γενικά αυτόματα από το PPTA αλλά λιγότερο γενικά από το HPA, αποκλείεται από την αναζήτηση όταν χρησιμοποιείται το HPA. Στην επόμενη ενότητα θα επεξηγηθεί η συνολική ποιότητα των υποθέσεων σε αυτό το μικρότερο χώρο αναζήτησης.



Σχήμα 1. Το HPA που μοντελοποιεί τις συνόδους του πίνακα 1. Οι αριθμοί μέσα σε κάθε κατάσταση δηλώνουν το πλήθος των αιτήσεων για την αντίστοιχη σελίδα και το πλήθος των φορών που αυτή ήταν τελευταία στις συνόδους του συνόλου.

Είναι λοιπόν ενδιαφέρον να συγκριθεί το αποτέλεσμα της εκμάθησης γραμματικών ξεκινώντας από ένα HPA, αντί για ένα PPTA. Ωστόσο, η εφαρμογή των υπαρχόντων αλγορίθμων εκμάθησης στοχαστικών κανονικών γραμματικών, οι οποίοι είναι σχεδιασμένοι για το PPTA δεν είναι πάντα εφικτή. Αυτό οφείλεται σε παραδοχές για τη δομή του αυτομάτου σε κάθε βήμα της διαδικασίας εκμάθησης. Για παράδειγμα, ο αλγόριθμος Alergia, ο οποίος φαίνεται στον πίνακα 2, αποφεύγει το ενδεχόμενο της άπειρης αναδρομής προσδιορίζοντας μια ακολουθία συγχωνεύσεων η οποία εξασφαλίζει ότι μία από τις καταστάσεις είναι πάντα η ρίζα ενός υποδέντρου στο PPTA. Επομένως, ο Alergia δεν μπορεί να εφαρμοστεί αυτούσιος σε ένα αυθαίρετο αυτόματο. Ωστόσο, το HPA έχει μια ενδιαφέρουσα ιδιότητα η οποία επιτρέπει την εφαρμογή του αλγορίθμου παρά την δομή του αυτομάτου. Όπως φαίνεται και στο σχήμα 1, κάθε σύμβολο A_i που μαρκάρει μια μετάβαση καθορίζει πλήρως την κατάσταση προορισμό, ανεξάρτητα από την τρέχουσα κατάσταση. Η πρώτη συνέπεια αυτής της ιδιότητας είναι ότι η συγχώνευση δεν μπορεί να προκαλέσει μη ντετερμινισμό. Μη ντετερμινιστικά αυτόματα μπορούν εν γένει να προκύψουν όταν συγχωνεύονται δύο καταστάσεις οι οποίες έχουν μεταβάσεις με το ίδιο σύμβολο s προς διαφορετικές καταστάσεις. Η συγχωνευμένη κατάσταση θα είναι τότε μη ντετερμινιστική επειδή θα υπάρχουν δύο καταστάσεις στις οποίες θα μπορεί να οδηγήσει το s. Για να αρθεί ο μη ντετερμινισμός, οι καταστάσεις που τον προκαλούν συγχωνεύονται αναδρομικά. Στην περίπτωση του HPA, οι μεταβάσεις με το ίδιο σύμβολο οδηγούν πάντα στην ίδια κατάσταση, οπότε δεν μπορεί να προκληθεί μη ντετερμινισμός. Το άλλο αποτέλεσμα της παραπάνω ιδιότητας είναι ότι ο έλεγχος συμβατότητας δύο καταστάσεων απλουστεύεται πολύ. Για να ελεγχθεί η συμβατότητα δύο καταστάσεων δεν είναι πλέον αναγκαίος ο έλεγχος της συμβατότητας όλων των επόμενων καταστάσεων, καθώς αυτές είναι πια ταυτόσημες και επομένως είναι εγγυημένα συμβατές.

Πίνακας 2. Ο αλγόριθμος ALERGIA. Οι φορές που συναντήθηκε η κατάσταση k κατά την ανάλυση του δείγματος δηλώνεται με $C(k)$. Με $C(k,\#)$ δηλώνονται οι φορές που η κατάσταση k ήταν τελευταία σε μια σύνοδο. Το $C(k,x)$ είναι οι φορές που το τερματικό σύμβολο x συναντήθηκε στο δείγμα όταν το αυτόματο ήταν στην κατάσταση k. Το $\delta(k,x)$ δηλώνει την επόμενη κατάσταση της k όταν συναντάται το x.

```
ALERGIA(S, α)
  A = PPTA(S)
  for j = 2 to |A| do
    for i = 1 to j − 1 do
      if COMPATIBLE(i, j, α)
        MERGE(i, j)
        DETERMINIZE(A)
---
COMPATIBLE(i, j, α)
  if DIFFERENT(C(i), C(j), C(i, #), C(j, #), α)
    return false
  for each terminal t do
    if DIFFERENT(C(i), C(j), C(i, t), C(j, t), α)
      return false
    if not COMPATIBLE(δ(i, t), δ(j, t), α)
      return false
  return true
---
DIFFERENT(n_i, n_j, f_i, f_j, α)
```

return $\left|\frac{f_i}{n_i} - \frac{f_j}{n_j}\right| > \sqrt{\frac{1}{2}\ln\frac{2}{\alpha}}\left(\frac{1}{\sqrt{n_i}} + \frac{1}{\sqrt{n_j}}\right)$

3 Πειράματα

Για την αξιολόγηση της μεθόδου μας χρησιμοποιήθηκαν δύο σύνολα δεδομένων. Το πρώτο σύνολο ήταν το MSWeb[1] και το δεύτερο ήταν δεδομένα χρήσης από το δικτυακό τόπο "Information Retrieval in Chemistry"[2], που φιλοξενείται στο ΕΚΕΦΕ «Δημόκριτοσ». Τα δεδομένα που χρησιμοποιήθηκαν στην πειραματική αξιολόγηση της δουλειάς μας ήταν ήδη χωρισμένα σε συνόδους και είχε ήδη γίνει κάποια αφαίρεση θορύβου [14], αφαιρώντας για παράδειγμα συνόδους που οφείλονταν σε spiders. Αυτό εμπόδισε οποιοδήποτε πειραματισμό με διάφορους χρόνους λήξης των συνόδων. Μια πρόσφατη δουλειά που μελετά το θέμα του χρόνου λήξης μιας συνόδου βρίσκεται στο [12]. Για το διαχωρισμό μεταξύ διαφορετικών συνόδων του ίδιου χρήστη, οι σύνοδοι του δεύτερου συνόλου παράχθηκαν περιορίζοντας το χρόνο μεταξύ δύο συνεχόμενων αιτήσεων από την ίδια IP διεύθυνση στα εξήντα λεπτά. Αν ο χρόνος μεταξύ δύο συνεχόμενων αιτήσεων υπερέβαινε τα εξήντα λεπτά, η δεύτερη αίτηση θεωρούταν η αρχή μιας νέας συνόδου. Τα σύνολο MSWeb ήταν χωρισμένο σύμφωνα με την ταυτότητα του χρήστη, αντί απλώς με την IP διεύθυνση. Επομένως ο περιορισμός του χρόνου μεταξύ διαδοχικών αιτήσεων δεν ήταν απαραίτητος. Το σύνολο MSWeb περιείχε περίπου 38000 συνόδους ενώ το σύνολο της χημείας 4500 συνόδους. Για τα πειράματα, τα δύο τρίτα των συνόδων χρησιμο-

[1] Δημόσια διαθέσιμο στο http://www.ics.uci.edu/∼mlearn/MLRepository.html

[2] http://macedonia.chem.demokritos.gr

ποιήθηκαν για την εκμάθηση του αυτομάτου και οι υπόλοιπες χρησιμοποιήθηκαν για την αξιολόγησή του.

Τα προβλήματα στα οποία τα αυτόματα αποτιμήθηκαν σχεδιάστηκαν ώστε να συλλάβουν την ικανότητα τους στην πρόβλεψη της συμπεριφοράς του χρήστη και την ικανότητα τους να παρέχουν καλές προτάσεις για την επόμενη σελίδα που θα ζητούσε ο χρήστης. Μετά τη φάση εκπαίδευσης, κρύβαμε την τελευταία σελίδα κάθε συνόδου στο σύνολο ελέγχου και δίναμε στο αυτόματο τις υπόλοιπες σελίδες κάθε συνόδου. Αυτές θα οδηγούσαν το αυτόματο σε μια συγκεκριμένη κατάσταση από την οποία μπορούμε στη συνέχεια να κατασκευάσουμε μια λίστα προτάσεων βασισμένοι στις πιο πιθανές μεταβάσεις από αυτή την κατάσταση. Μια επιτυχής πρόβλεψη συμβαίνει όταν η κρυμμένη σελίδα είναι ανάμεσα στις σελίδες που είναι πιο πιθανό να ζητηθούν από την τρέχουσα κατάσταση.

Για την αποτίμηση μιας λίστας προτάσεων προς το χρήστη χρησιμοποιήθηκε η μετρική της αναμενόμενης χρησιμότητας [6]. Η αναμενόμενη χρησιμότητα μιας προταθείσας σελίδας είναι απλώς η πιθανότητα να ζητηθεί επί τη χρησιμότητά της. Όπως και στο [6], η χρησιμότητα v_{aj} μιας σελίδας j για ένα χρήστη a είναι ένα αν η j ανήκει στις σελίδες της συνόδου και μηδέν διαφορετικά. Η πιθανότητα να ζητηθεί η i-οστή σελίδα σε μία λίστα επιλέχθηκε να φθίνει εκθετικά συναρτήσει του i. Έτσι η αναμενόμενη χρησιμότητα μιας λίστα σελίδων (ταξινομημένης σε φθίνουσα σειρά πιθανότητας μετάβασης) είναι:

$$R_a = \sum_{j=0}^{l} \frac{v_{aj}}{2^{j/h}}$$

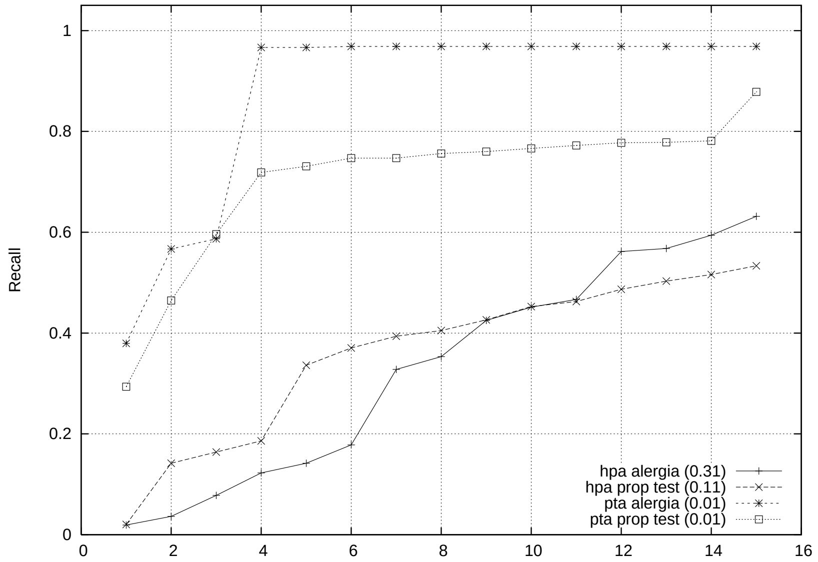
όπου l είναι ο αύξων αριθμός της τελευταίας σελίδας στη λίστα και h είναι ο χρόνος ημισείας ζωής. Παρόμοια με άλλα φαινόμενα εκθετικής αποσύνθεσης, ο χρόνος ημισείας ζωής είναι ο αύξων αριθμός της σελίδας η οποία έχει πιθανότητα 0.5 να ζητηθεί. Στα πειράματα μας, θέσαμε την πέμπτη σελίδα να έχει αυτή την πιθανότητα. Τέλος, για ένα σύνολο από συνόδους χρηστών με τις αντίστοιχες λίστες προτάσεων, η αναμενόμενη χρησιμότητα είναι

$$R = 100 \frac{\sum_a R_a}{\sum_a R_a^{max}}$$

όπου το R_a^{max} είναι η μέγιστη δυνατή χρησιμότητα αν η κορυφή της λίστας περιείχε μόνο σελίδες με v_{aj} ίσο με ένα.

Η ακρίβεια πρόβλεψης ενός αυτομάτου μετράται χρησιμοποιώντας την ανάκληση της λίστας που προτείνει. Η ανάκληση ενός συστήματος ανάκτησης πληροφοριών ορίζεται ως το ποσοστό των συνολικά διαθέσιμων σχετικών εγγράφων που ανακτά. Σε αυτή την περίπτωση η κρυμμένη σελίδα θεωρείται ότι είναι σχετική. Για τον υπολογισμό της ανάκλησης, υπολογίζουμε το ποσοστό των συνόδων στις οποίες η κρυμμένη σελίδα ανήκει στην λίστα που προτείνεται ως συνάρτηση του μήκους αυτής της λίστας. Η γραφική παράσταση αυτής της συνάρτησης υποδεικνύει ένα κατάλληλο μέγεθος της λίστας των σελίδων που πρέπει να προταθούν στο χρήστη.

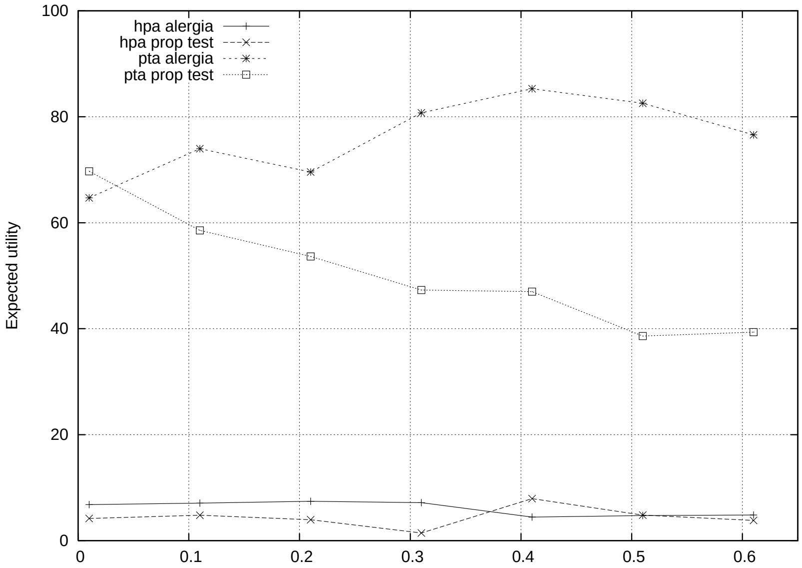
Στο σχήμα 2 φαίνεται η ανάκληση των καλύτερων αυτομάτων για τα δεδομένα της χημείας. Τα αντίστοιχα αποτελέσματα για τα δεδομένα του MSWeb είναι παρόμοια [11] και γι' αυτό παραλείπονται. Το υπόμνημα δείχνει για κάθε καμπύλη την αρχική υπόθεση που χρησιμοποιήθηκε, τον αλγόριθμο που εφαρμόστηκε και την επιλεγμένη τιμή της παραμέτρου α. Με τον όρο "prop test" εννοούμε τον αλγόριθμο ALERGIA με το στατιστικό έλεγχο που προτείνεται στο [10]. Από την άλλη ο όρος "alergia" αναφέρεται στον αρχικό αλγόριθμο που χρησιμοποιεί το φράγμα Hoeffding.



Σχήμα 2. Η ανάκληση ορισμένων αυτομάτων που προέκυψαν από το σύνολο της χημείας. Ο άξονας x αναπαριστά το μήκος της προτεινόμενης λίστας

Από αυτή τη γραφική παράσταση φαίνεται ότι τα μοντέλα που προέκυψαν με το PPTA ως αρχική υπόθεση, είναι σαφώς ανώτερα των μοντέλων που προέκυψαν από τη Μαρκοβιανή αλυσίδα. Η γραφική παράσταση δείχνει επίσης ότι οι μέθοδοι εκμάθησης γραμματικών μπορούν να παράγουν μοντέλα με περισσότερο από 90% ανάκληση στις πρώτες 4 σελίδες της λίστας προτάσεων. Από την άλλη πλευρά, τα αυτόματα που προκύπτουν από Μαρκοβιανές αλυσίδες έχουν μικρή ανάκληση η οποία αυξάνεται σταθερά με το μήκος της λίστας. Αυτό σημαίνει ότι οι προτάσεις τους είναι στην πραγματικότητα εικασίες της συμπεριφοράς του χρήστη.

Η αναμενόμενη χρησιμότητα των αυτομάτων που προκύπτουν για τις διαφορετικές τιμές της παραμέτρου α για τα δεδομένα του MSWeb φαίνεται στο σχήμα 3. Τα αντίστοιχα αποτελέσματα για τα δεδομένα της χημείας είναι παρόμοια [11] και γι΄ αυτό παραλείπονται. Παρατηρώντας τη γραφική παράσταση, αποφασίστηκε να τεθεί το μέγιστο μήκος της λίστας προτάσεων σε δέκα προτάσεις. Καθώς το α αυξάνεται λιγότερες καταστάσεις συγχωνεύονται και το αυτόματο είναι λιγότερο γενικό. Αυτό σημαίνει ότι η αναμενόμενη χρησιμότητα θα πρέπει να ελαττώνεται καθώς το α αυξάνεται. Και πάλι τα αυτόματα που προέρχονται από Μαρκοβιανές αλυσίδες δίνουν κατώτερα μοντέλα συγκρινόμενα με τα αυτόματα που προκύπτουν από τους συμβατικούς αλγορίθμους. Ωστόσο, η αναμενόμενη χρησιμότητα των αυτομάτων που προέρχονται από Μαρκοβιανές αλυσίδες φαίνεται να είναι λίγο μεγαλύτερη από αυτή που θα έδινε η αρχική Μαρκοβιανή αλυσίδα. Αυτό το συμπέρασμα εξάγεται επειδή για α ίσο με ένα παίρνουμε την αρχική υπόθεση, καθώς δεν θα γίνει καμία συγχώνευση, και επειδή η αναμενόμενη χρησιμότητα μειώνεται καθώς το α αυξάνεται. Ωστόσο, η αναμενόμενη χρησιμότητα στην περίπτωση αυτών των αυτομάτων είναι πολύ χαμηλότερη από εκείνη των αυτομάτων που προκύπτουν από PPTA και είναι επίσης πολύ ευαίσθητη στις μεταβολές του α. Η τελευταία παρατήρηση είναι ένδειξη ότι τα καλά μοντέλα έχουν αποκλειστεί από το χώρο αναζήτησης.



Σχήμα 3. Η αναμενόμενη χρησιμότητα ως συνάρτηση του α για το σύνολο MSWeb. Το μέγιστο μήκος της προτεινόμενης λίστας είχε τεθεί σε 10.

4 Συμπεράσματα

Αποτιμήσαμε δύο μοντέλα για την συμπεριφορά πλοήγησης των χρηστών σε ένα δικτυακό τόπο. Και τα δύο μπορούν να ειδωθούν ως στοχαστικές κανονικές γραμματικές, αλλά το ένα προκύπτει από το παραδοσιακό δενδρικό αυτόματο προθεμάτων ενώ το άλλο από μια Μαρκοβιανή αλυσίδα πρώτης τάξης η οποία έχει χρησιμοποιηθεί παλαιότερα για τη μοντελοποίηση αυτής της συμπεριφοράς. Σε αυτή τη δουλειά παρουσιάστηκαν τα αποτελέσματα του γνωστού αλγόριθμου Alergia κατάλληλα προσαρμοσμένου ώστε να δουλεύει και σε Μαρκοβιανές αλυσίδες ως αρχικές υποθέσεις. Δύο στατιστικοί έλεγχοι χρησιμοποιήθηκαν για να αποφασιστεί αν δύο καταστάσεις θα έπρεπε να συγχωνευθούν, ο έλεγχος Hoeffding και ο έλεγχος αναλογιών. Η διαφορά μεταξύ των δύο ελέγχων είναι ότι ο δεύτερος είναι αυστηρότερος του πρώτου, οδηγώντας σε λιγότερες συγχωνεύσεις καταστάσεων και μεγαλύτερα αυτόματα. Αυτό έχει δύο αποτελέσματα. Από τη μια πλευρά ο έλεγχος Hoeffding οδηγεί σε πιο γενικά αυτόματα, κάτι που διαφαίνεται στην υψηλότερη ανάκληση και αναμενόμενη χρησιμότητα που επιτυγχάνουν. Από την άλλη, ο έλεγχος αναλογιών είναι λιγότερο πιθανό να κάνει μια κακή συγχώνευση, λόγω της αναπόφευκτης παρουσίας θορύβου στα δεδομένα. Όταν συγχωνεύονται δύο καταστάσεις που δεν είναι συμβατές, το προκύπτον αυτόματο μπορεί να είναι πολύ διαφορετικό από το στόχο. Αυτό μπορεί να οδηγήσει σε αποτελέσματα τα οποία είναι πολύ ευαίσθητα στις τιμές του α. Αυτός είναι ο λόγος που στο σχήμα 3 η αναμενόμενη χρησιμότητα φθίνει στην περίπτωση του ελέγχου αναλογιών. Στην περίπτωση του ελέγχου Hoeffding ωστόσο, η καμπύλη είναι πολύ ευαίσθητη στην τιμή του α.

Χρησιμοποιήσαμε δεδομένα από δύο πολύ διαφορετικούς δικτυακούς τόπους. Και στα δύο σύνολα δεδομένων τα αυτόματα που προέκυψαν από το δενδρικό αυτόματο προθεμάτων είχαν καλύτερη απόδοση όσον αφορά τις μετρικές της ανάκλησης και της αναμενόμενης χρησιμότητας. Αυτό υποδηλώνει ότι, αν και οι στοχαστικές κανονικές γλώσσες δεν είχαν προηγουμένως χρησιμοποιηθεί στη μοντελοποίηση της πλοήγησης των χρηστών στον Παγκόσμιο Ιστό, είναι πιο κατάλληλες από τις Μαρκοβιανές αλυσίδες που έχουν χρησιμοποιηθεί συχνά για αυτό το σκοπό. Αυτό αποδίδεται στην εκφραστικότητα των στοχαστικών κανονικών γλωσσών.

Εκτός από την περαιτέρω εμπειρική αξιολόγηση, ενδιαφερόμαστε και για τη χρήση των αυτόματων που προκύπτουν με διαφορετικούς τρόπους. Οι συγχωνευμένες καταστάσεις είναι ουσιαστικά σύνολα ιστοσελίδων που θα μπορούσαν να ειδωθούν ως συσταδοποίηση των σελίδων ενός δικτυακού τόπου. Τα προκαταρκτικά πειράματα με το σύνολο MSWeb είναι ελπιδοφόρα δεδομένου ότι έχουν δείξει την δημιουργία συστάδων με φυσικό νόημα. Επίσης εξετάζονται διαφορετικά κριτήρια με τα οποία οι καταστάσεις θα μπορούσαν να συγχωνευθούν. Εφόσον η συγχώνευση των ιστοσελίδων σύμφωνα με τις πιθανότητες μετάβασής τους φαίνεται αρχικά περίεργη, προσπαθούμε να υιοθετήσουμε κριτήρια που θα παράσχουν ένα πιο διαισθητικό νόημα σε μία συγχωνευμένη κατάσταση. Τέλος, σχεδιάζουμε να συγκρίνουμε τα αποτελέσματά μας με εκείνα γνωστών αλγορίθμων ανακάλυψης ακολουθιακών προτύπων [1,9].

Ευχαριστίες

Θα θέλαμε να ευχαριστήσουμε τα μέλη της ομάδας «Ανάκτηση πληροφοριών στη χημεία»του ΕΚΕΦΕ «Δημόκριτος»(Ε. Βαρβέρη, Α. Βαρβέρης και Π. Τελώνης) για την παροχή των δεδομένων της χημείας. Προσωπικά, θα ήθελα να ευχαριστήσω τον υποψήφιο διδάκτορα Δημήτριο Πιερράκο για την υποστήριξη που μου παρείχε κατά την εκπόνηση αυτής της εργασίας.

Αναφορές

1. R. Agrawal and R. Srikant. *Mining sequential patterns.* In Proceedings of the Eleventh International Conference on Data Engineering, 1995.
2. D. Albrecht, I. Zukerman and A. Nicholson. *Pre-sending documents on the WWW: A comparative study.* Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, pp. 1274-1279, 1999.
3. C. Anderson, P. Domingos, and D. Weld. *Adaptive Web Navigation for Wireless Devices.* Proceedings of the 17th IJCAI, 2001.
4. J. Borges and M. Levene. *Data Mining of User Navigation Patterns.* Lecture Notes in Computer Science, Vol. 1836, pp. 92-111, 1999.
5. J. Borges. *A Data Mining Model to Capture User Web Navigation Patterns.* PhD dissertation, 2000.
6. J. S. Breese, D. Heckerman and C. Kadie. *Empirical Analysis of Predictive Algorithms for Collaborative Filtering.* Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, 1998.
7. I. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White. *Visualization of Navigation Patterns on a Web Site Using Model Based Clustering.* Technical Report MSR-TR-00-18, 2000.
8. R. Carrasco and J. Oncina. *Learning Regular Grammars by Means of a State Merging Method.* Proceedings of the ICGI, 1994.
9. R. Cooley, B. Mobasher, and J. Srivastava. *Web Mining: Information and Pattern Discovery on the World Wide Web.* In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, 1997.
10. A. Habrard, M. Bernard and M. Sebban. *Improvement of the State Merging Rule on Noisy Data in Probabilistic Grammatical Inference.* Proceedings of the Fourteenth ECML, 2003.
11. N. Karampatziakis, G. Paliouras, D. Pierrakos and P. Stamatopoulos. *Navigation Pattern Discovery Using Grammatical Inference.* Proceedings of the ICGI, 2004.
12. R. Kohavi and R. Parekh. *Ten supplementary analyses to Improve E-commerce Web Sites.* In Proceedings of the Fifth WEBKDD workshop, 2003.
13. S.H. Muggleton, C. H. Bryant and A. Srinivasan. *Learning Chomsky-like grammars for biological sequence families.* In Proceedings of the Seventeenth International Conference on Machine Learning, 2000.
14. G. Paliouras, C. Papatheodorou, V. Karkaletsis and C. D. Spyropoulos. *Clustering the Users of Large Web Sites into Communities.* Proceedings of International Conference on Machine Learning (ICML), 2000.
15. R. Parekh and V. Honavar. *Grammar Inference, Automata Induction and Language Acquisition.* Invited chapter in Handbook of Natural Language Processing. R. Dale, H. Moisl, and H. Somers, editors. Marcel Dekker, New York, 2000.
16. D. Pierrakos, G. Paliouras, C. Papatheodorou and C. D. Spyropoulos. *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.
17. J. Pitkow and P. Pirolli. *Mining longest repeating subsequences to predict WWW surfing.* Proceedings of the 1999 USENIX Annual Technical Conference, 1999.
18. M. Spiliopoulou, L. C. Faulstich, and K. Wilkler. *A data miner analyzing the navigational behavior of Web users.* In Proceedings of the Workshop on Machine Learning in User Modeling of the ACAI99, 1999.
19. J. Srivastava, R. Cooley, M. Deshpande, and P. T. Tan. *Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data.* SIGKDD Explorations, 1(2), pp. 12-23, 2000.