Οδηγός Χρήσης του ΔΕiΧΤο

ΔΕiΧΤο V2.8.8.0

1 Μαρτίου 2008

Κώστας Ντονάς
kntonas@gmail.com

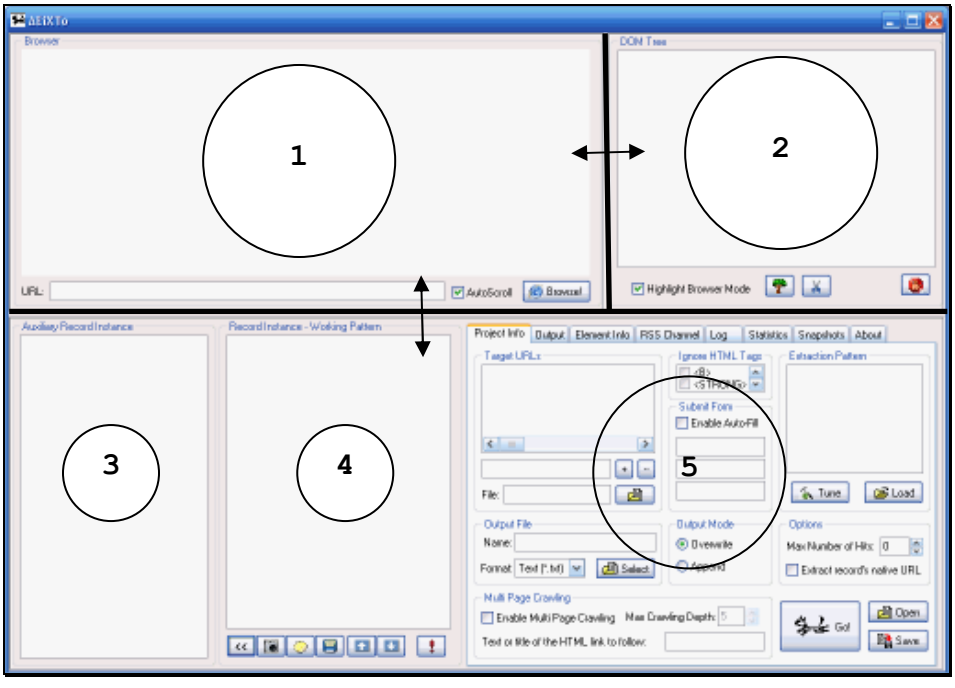
Πίνακας Περιεχομένων

Πίνακας Εικόνων

ΔEiXTo

Το *ΔEiXTo* (ή αλλιώς *DEiXTo*) είναι ένα ισχυρό εργαλείο εξαγωγής περιεχομένου από ιστοσελίδες (web data extraction tool). Έχει να επιδείξει αρκετά προηγμένα χαρακτηριστικά και πολύ υψηλά επίπεδα ακρίβειας, τουλάχιστον για την πλειοψηφία των περιπτώσεων. Το εγχειρίδιο αυτό παρουσιάζει αναλυτικά τη λειτουργικότητα του ΔEiXTo, το όνομα του οποίου προέκυψε από τα αρχικά των λέξεων *Data Extraction Tool* και την ελληνική φράση «δείξ'το».

Το εργαλείο παρέχει ένα εύχρηστο γραφικό περιβάλλον (GUI) μέσω του οποίου ο χρήστης μπορεί σχετικά εύκολα να κατασκευάσει αποτελεσματικούς κανόνες εξαγωγής (extraction rules ή wrappers) και να τους εκτελέσει ώστε τελικά να εξάγει τα επιθυμητά δεδομένα από ιστοσελίδες που τον ενδιαφέρουν. Στην Εικόνα 1 απεικονίζεται το γραφικό περιβάλλον του ΔEiXTo και απαριθμούνται τα σημαντικότερα επιμέρους τμήματα του παραθύρου της εφαρμογής, τα οποία θα αναλυθούν στη συνέχεια. Να σημειωθεί πως μέσω ενός οριζόντιου και ενός κατακόρυφου splitter, ο χρήστης μπορεί να μεταβάλλει το μέγεθος ορισμένων τμημάτων του εργαλείου.



Εικόνα 1: Το γραφικό περιβάλλον του ΔEiXTo

Ενσωματωμένο Πρόγραμμα Πλοήγησης

Στο ΔΕiΧΤο, κυρίαρχο ρόλο παίζει το ενσωματωμένο πρόγραμμα πλοήγησης Ιστού το οποίο βρίσκεται στο πάνω αριστερό τμήμα του παραθύρου της εφαρμογής (Εικόνα 1, Περιοχή 1). Αν λοιπόν ο χρήστης ενδιαφέρεται να εξάγει δεδομένα από μια συγκεκριμένη HTML ιστοσελίδα ή ένα δικτυακό τόπο, η πρώτη ενέργεια που έχει να κάνει είναι να πληκτρολογήσει το σχετικό *URL* στη γραμμή διευθύνσεων και να πατήσει το κουμπί πλοήγησης (`Browse`). Να σημειωθεί πως το πρόγραμμα πλοήγησης μπορεί να ανακτήσει και HTML έγγραφα τα οποία είναι αποθηκευμένα τοπικά στο δίσκο του χρήστη μέσω του σχήματος *file://path*. Σε περίπτωση βέβαια που η λήψη μιας σελίδας για κάποιο λόγο αποτύχει ή περάσει χρονικό διάστημα ίσο με το μέγιστο επιτρεπόμενο χρόνο αναμονής (timeout), ο οποίος έχει οριστεί στα 5 λεπτά, τότε γίνονται οι κατάλληλες ενέργειες και εμφανίζονται σχετικά μηνύματα. Επίσης, με ‘Alt+Αριστερό Βέλος’ και ‘Alt+Δεξιό Βέλος’ ο χρήστης μπορεί να «πηγαίνει» Back και Forward αντίστοιχα.

Πρέπει ωστόσο να σημειωθεί πως το εργαλείο *δεν* έχει τη δυνατότητα να χειρίζεται HTML ιστοσελίδες που έχουν *πλαίσια* (frames). Αυτό οφείλεται στο γεγονός ότι η ύπαρξη πλαισίων σε μία σελίδα κάνει ιδιαίτερα πολύπλοκο το χειρισμό της και απαιτεί προγραμματιστικά ειδική αντιμετώπιση, καθώς κάθε πλαίσιο αποτελεί μία ξεχωριστή σελίδα. Βέβαια, η συντριπτική πλειοψηφία των ιστοσελίδων δεν περιέχει πλαίσια. Αυτό συνεπάγεται ότι η αξία του εργαλείου δε μειώνεται, τουλάχιστον σοβαρά, εξαιτίας της μη αντιμετώπισης τέτοιων περιπτώσεων.

myDOM Αναπαράσταση

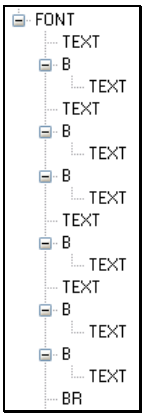
Η λειτουργία του ΔΕiΧΤο βασίζεται στο *Μοντέλο Αντικειμένου Εγγράφου* (Document Object Model ή DOM) της Κοινοπραξίας του Παγκόσμιου Ιστού (W3C). Το DOM συνιστά ένα αντικειμενοστραφές μοντέλο περιγραφής εγγράφων Ιστού και παρέχει μια δενδροειδή αναπαράσταση (tree representation) των ιστοσελίδων. Κρίθηκε λοιπόν απαραίτητη η οπτικοποίηση της DOM αναπαράστασης της προβαλλόμενης ιστοσελίδας και η αποτύπωση της σε ορατό συστατικό στο πάνω δεξιό μέρος του παραθύρου της εφαρμογής (Εικόνα 1, περιοχή 2). Να σημειωθεί πως από εδώ και στο εξής η δενδροειδής αυτή αναπαράσταση θα αναφέρεται ως *myDOM*.

Η κατασκευή της myDOM δενδρικής δομής μίας σελίδας πραγματοποιείται όταν ολοκληρώνεται η λήψη της σελίδας από το πρόγραμμα πλοήγησης και γίνεται

με χρήση των διεπαφών που προσφέρει το DOM και κλασικού αναδρομικού αλγορίθμου σάρωσης δέντρου κατά βάθος (depth first). Μάλιστα, για κάθε κόμβο του myDOM αποθηκεύονται χρήσιμες πληροφορίες. Τα δεδομένα που μπορούν να εξαχθούν είναι: για υπερσυνδέσμους (<A>) το `href` γνώρισμα τους, για εικόνες (<IMG>) το `src` γνώρισμα τους, για κόμβους κειμένου (TEXT) το κείμενο που περιέχουν, για FORM και INPUT στοιχεία (elements) το `name` γνώρισμα τους και για τα υπόλοιπα HTML στοιχεία το εσωκλειόμενο κείμενο τους (inner text). Επίσης, υπάρχει δυνατότητα εξαγωγής του πηγαίου κώδικα ενός HTML στοιχείου.

Απλοποίηση myDOM Αναπαράστασης

Ιδιαίτερα βοηθητική λειτουργία αποδείχτηκε η δυνατότητα να αγνοούνται κάποιοι τύποι HTML κόμβων (tag filtering) κατά την κατασκευή της myDOM αναπαράστασης μιας ιστοσελίδας, χωρίς όμως να αγνοείται και το εσωκλειόμενο κείμενο τους. Η λειτουργία αυτή υλοποιήθηκε διότι διαπιστώθηκε σε ορισμένες περιπτώσεις ότι κάποιοι τύποι στοιχείων δημιουργούσαν πρόβλημα τόσο στον εντοπισμό των στιγμιότυπων επιθυμητής πληροφορίας όσο και στην εξαγωγή των χρήσιμων δεδομένων. Για να γίνει το παραπάνω πιο σαφές ακολουθεί ένα χαρακτηριστικό παράδειγμα. Σε μία τυπική σελίδα αποτελεσμάτων του Google παρουσιάζεται σοβαρό πρόβλημα λόγω της έντονης (bold) γραφής κάποιων λέξεων. Τα <B> HTML στοιχεία αναγκάζουν ένα τμήμα κειμένου να «σπάσει» σε πολλά επιμέρους τμήματα, με συνέπεια τα στιγμιότυπα να περιλαμβάνουν δομές της μορφής που φαίνεται στην Εικόνα 2.



Εικόνα 2: Προβληματικό υποδένδρο λόγω ετικετών <B>

Η παραπάνω δομή δυσχεραίνει σημαντικά τόσο την εξαγωγή του ενιαίου κειμένου όσο και τον εντοπισμό των στιγμιότυπων καθώς το πλήθος των λέξεων σε

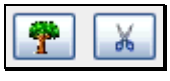
έντονη γραφή σε ένα στιγμιότυπο είναι μεταβλητό. Ωστόσο, με τη λειτουργία απλοποίησης (simplification) της myDOM αναπαράστασης, έπειτα από σχετική εντολή του χρήστη, τα στοιχεία που δημιουργούν το πρόβλημα αγνοούνται και το εσωτερικό τους κείμενο συγχωνεύεται με γειτονικούς κόμβους κειμένου. Έτσι, το παραπάνω υποδένδρο έπειτα από παράλειψη των <B> κατά την κατασκευή της myDOM δενδροειδούς αναπαράστασης μετατρέπεται σε αυτό που φαίνεται στην Εικόνα 3.

Εικόνα 3: Υποδένδρο μετά την απλοποίηση

Σαν αποτέλεσμα διευκολύνεται ο εντοπισμός όλων των στιγμιότυπων, αφού πλέον δεν υπάρχει εξάρτηση από τον αριθμό των <B> στοιχείων που έχει κάθε στιγμιότυπο και επιπρόσθετα είναι δυνατή η εξαγωγή του ενιαίου κειμένου. Είναι προφανής λοιπόν η χρησιμότητα και σημαντικά τα πλεονεκτήματα της λειτουργίας απλοποίησης της myDOM αναπαράστασης. Για την πραγματοποίηση της, πρέπει πρώτα ο χρήστης να επιλέξει τα στοιχεία που θέλει να απομακρύνει, από μία προκαθορισμένη λίστα HTML στοιχείων (Εικόνα 4) στην καρτέλα *Project Info* στην περιοχή 5 της εφαρμογής (Εικόνα 1).

Εικόνα 4: Λίστα με ετικέτες στοιχείων προς απομάκρυνση

Στη συνέχεια πρέπει ο χρήστης να πατήσει το κουμπί «κλαδέματος» του myDOM (Εικόνα 5), που βρίσκεται στην περιοχή 2 του παραθύρου της εφαρμογής. Ο χρήστης έχει επίσης τη δυνατότητα, εφόσον το επιθυμεί, να κατασκευάσει ξανά την πρωτότυπη myDOM αναπαράσταση, επιλέγοντας το σχετικό κουμπί, που είναι ακριβώς δίπλα σε αυτό της απλοποίησης (Εικόνα 5).



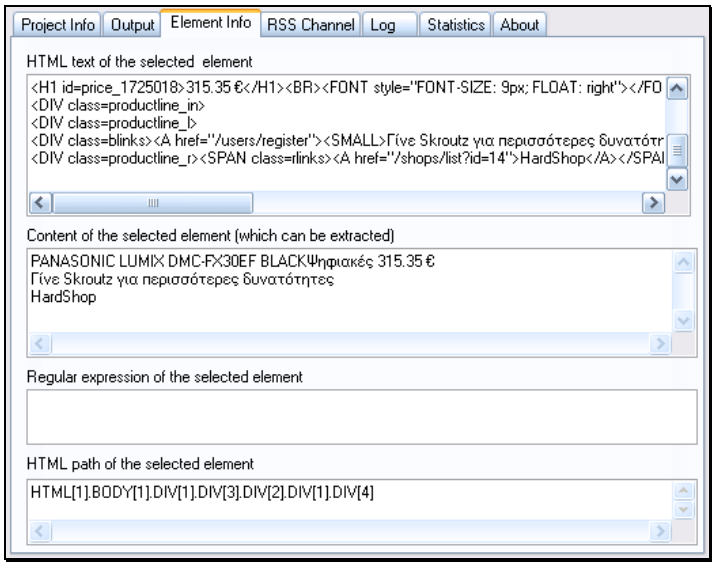
Εικόνα 5: Κουμπιά ανακατασκευής και κλαδέματος της myDOM αναπαράστασης

Δημιουργία Κανόνα Εξαγωγής

Όταν και εφόσον ολοκληρωθεί επιτυχώς η λήψη μιας ιστοσελίδας και γίνει η προβολή της μέσα στο πρόγραμμα πλοήγησης, πρέπει ο χρήστης να περιγράψει τη μορφή της επιθυμητής πληροφορίας, δηλαδή να κατασκευάσει ένα κανόνα εξαγωγής. Για το σκοπό αυτό, το πρόγραμμα πλοήγησης εμπλουτίστηκε με μία ελεγχόμενη κατάσταση λειτουργίας επισήμανσης (highlight mode), ώστε περιοχές της ιστοσελίδας που αντιστοιχούν σε ορατά HTML στοιχεία να χρωματίζονται διαφορετικά όταν ο κέρσορας περνά από πάνω τους. Έτσι, όταν ο κέρσορας είναι πάνω από το HTML έγγραφο και το σχετικό κουμπί ελέγχου (checkbox) στην περιοχή 2 της εφαρμογής είναι ενεργοποιημένο, τότε γίνεται επισήμανση του HTML στοιχείου στο οποίο αντιστοιχεί η θέση του mouse, εφόσον βέβαια αυτό είναι δυνατό. Η Εικόνα 6 είναι ενδεικτική αυτής της λειτουργίας.

Εικόνα 6: Λειτουργία οπτικής επισήμανσης

Επιπλέον, στην καρτέλα *Element Info*, στην περιοχή 5 στην Εικόνα 1, εμφανίζονται διάφορες χρήσιμες πληροφορίες για το επιλεγμένο στοιχείο (Εικόνα 7), όπως ο πηγαίος κώδικας του στοιχείου (outer HTML), τα δεδομένα που μπορούν να εξαχθούν από αυτόν και η απόλυτη διαδρομή θέσης του μέσα στο έγγραφο.

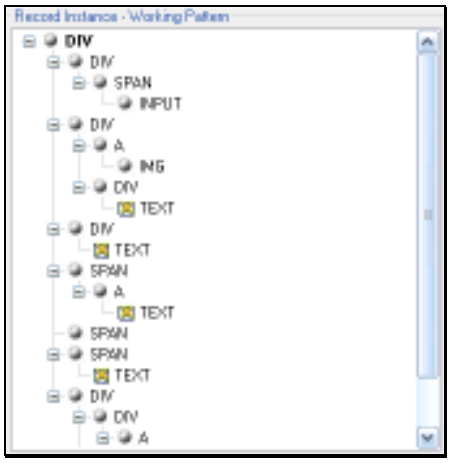


Εικόνα 7: Βοηθητικές πληροφορίες για το επιλεγμένο myDOM στοιχείο

Ο χρήστης μπορεί εύκολα και γρήγορα να κατασκευάσει στιγμιότυπο της επιθυμητής πληροφορίας με επιλογή της σχετικής λειτουργίας στο τοπικό μενού (popup menu) του HTML στοιχείου που τον ενδιαφέρει. Τότε, δημιουργείται δενδρικός κανόνας, ο οποίος φαίνεται στο πλαίσιο της περιοχής 4 του παραθύρου της εφαρμογής (Εικόνα 1). Η δενδρική αυτή δομή είναι το υποδένδρο της myDOM αναπαράστασης της ιστοσελίδας με ρίζα τον κόμβο που υπέδειξε ο χρήστης.

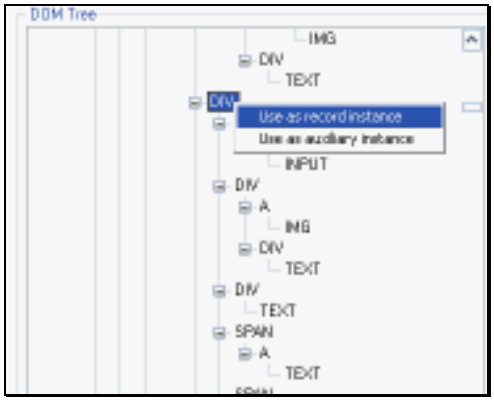
Στην Εικόνα 8 φαίνεται τμήμα του κανόνα που αντιστοιχεί στο στοιχείο που επισημαίνεται στην Εικόνα 6. Κάθε κόμβος της δομής έχει ένα όνομα, το οποίο είναι είτε η ετικέτα του αντίστοιχου HTML στοιχείου είτε `TEXT` αν πρόκειται για κόμβο κειμένου. Μάλιστα, ο ριζικός κόμβος του κανόνα φαίνεται με έντονα γράμματα. Πρέπει να σημειωθεί ότι ένας νέος κανόνας είναι ρυθμισμένος εκ των προτέρων, έτσι ώστε να εξάγει τα επιμέρους τμήματα κειμένου που εσωκλείονται στους `TEXT` κόμβους κάθε εντοπισμένου στιγμιότυπου.

Η δομή του κανόνα έχει διττό ρόλο. Είναι ταυτόχρονα το τρέχον πρότυπο (working pattern) που θα χρησιμοποιηθεί για την εξαγωγή πληροφορίας αλλά και ένα στιγμιότυπο της επιθυμητής πληροφορίας (record instance). Έτσι, όταν ο χρήστης επιλέγει ένα κόμβο του κανόνα, χρωματίζεται διαφορετικά η περιοχή της προβαλλόμενης ιστοσελίδας στην οποία αντιστοιχεί ο κόμβος αυτός. Αυτό διευκολύνει ιδιαίτερα τη διαδικασία κατάλληλης ρύθμισης του κανόνα, ώστε να μεγιστοποιηθεί η αποτελεσματικότητα του.



Εικόνα 8: Τμήμα δενδρικού κανόνα

Σε ορισμένες περιπτώσεις όμως δεν είναι δυνατή η οπτική επισήμανση ενός στοιχείου καθώς δεν είναι όλα τα στοιχεία άμεσα επιλέξιμα. Για παράδειγμα, τα μη ορατά στοιχεία δεν είναι δυνατό να επισημανθούν με υπόδειξη μέσω mouse. Τότε, ο χρήστης μπορεί να χρησιμοποιήσει τη myDOM αναπαράσταση για να κατασκευάσει τον κανόνα εξαγωγής. Αρκεί να επιλέξει στο τοπικό μενού του myDOM κόμβου τη λειτουργία δημιουργίας στιγμιότυπου επιθυμητής πληροφορίας. Τότε, δημιουργείται δενδρικός κανόνας, ο οποίος αποτελείται από το υποδένδρο του myDOM με ρίζα τον κόμβο αυτό. Η διαδικασία αυτή αποτυπώνεται στην Εικόνα 9.



Εικόνα 9: Δημιουργία κανόνα μέσω της myDOM αναπαράστασης

Επιπρόσθετα, υλοποιήθηκε κατάσταση *συγχρονισμένης* λειτουργίας μεταξύ προγράμματος πλοήγησης και myDOM αναπαράστασης, κατά την οποία η επιλογή ενός myDOM κόμβου ενεργοποιεί, εφόσον αυτό είναι δυνατό, την οπτική επισήμανση του αντίστοιχου στοιχείου της προβαλλόμενης σελίδας στο πρόγραμμα πλοήγησης και αντίστροφα.

Να σημειωθεί πως όταν ο κέρσορας βγαίνει εκτός προγράμματος πλοήγησης και το κουμπί ελέγχου επισήμανσης είναι ενεργό, τότε παύει και η οπτική επισήμανση κάποιου στοιχείου. Οπότε, για τις περιπτώσεις που δεν είναι εφικτή η επισήμανση του επιθυμητού στοιχείου μέσω mouse στο πρόγραμμα πλοήγησης, η συνήθης τακτική είναι να επισημαίνει ο χρήστης ένα στοιχείο που είναι κοντά σε αυτό που θέλει να «πιάσει», έπειτα να απενεργοποιεί τη λειτουργία επισήμανσης με επιλογή της αντίστοιχης λειτουργίας στο αναδυόμενο μενού και τέλος να υποδεικνύει στο myDOM δέντρο το στοιχείο που πραγματικά τον ενδιαφέρει.

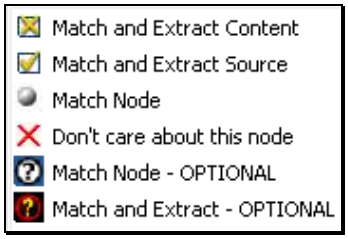
Η κατασκευή αποτελεσματικών κανόνων εξαγωγής προϋποθέτει την προσεκτική επιλογή αντιπροσωπευτικού παραδείγματος επιθυμητής πληροφορίας, το οποίο χρησιμεύει ως πρότυπο για την ανακάλυψη όλων των στιγμιότυπων επιθυμητών αντικειμένων που εμφανίζονται στη σελίδα. Το γεγονός που εκμεταλλεύεται το ΔΕiΧΤο είναι ότι τα σημασιολογικά σχετικά αντικείμενα έχουν κοινό ή παρόμοιο στυλ παρουσίασης καθώς και σχεδόν ίδια HTML δομή. Καλή πρακτική συνήθως αποτελεί η υπόδειξη της μικρότερης δυνατής περιοχής δεδομένων που περιέχει όλες τις πληροφορίες που ενδιαφέρουν το χρήστη σε ένα στιγμιότυπο. Σημαντικότατο πλεονέκτημα του συστήματος είναι η οπτικοποίηση της διαδικασίας, η οποία διευκολύνει πολύ την κατασκευή κανόνων εξαγωγής. Δεν είναι υπερβολή ότι πολλές φορές φτιάχνονται εύκολα σύνθετοι κανόνες μέσα σε λίγα δευτερόλεπτα.

Ρύθμιση Κανόνα Εξαγωγής

Σχεδόν πάντα, η απλή υπόδειξη από το χρήστη ενός στιγμιότυπου της επιθυμητής πληροφορίας δεν αρκεί για τον εντοπισμό και την εξαγωγή όλης, και μόνο αυτής, της επιθυμητής πληροφορίας. Αυτό συμβαίνει συχνά για παράδειγμα σε περιπτώσεις όπου εμφανίζονται πολλαπλά στιγμιότυπα επιθυμητής πληροφορίας στην ίδια ιστοσελίδα και τα οποία παρουσιάζουν μικρές ή και μεγαλύτερες παραλλαγές στη δομή τους. Οι παραλλαγές αυτές συνήθως οφείλονται σε προαιρετικά τμήματα δεδομένων (missing fields). Επίσης, ο χρήστης συνήθως επιθυμεί συγκεκριμένα πεδία δεδομένων του κάθε στιγμιότυπου και όχι το σύνολο των δεδομένων που αυτό περιέχει. Μάλιστα, σε αρκετές περιπτώσεις, μπορεί να ενδιαφέρει το χρήστη όχι ένα απλό τμήμα κειμένου αλλά κάποια γνωρίσματα ορισμένων κόμβων. Χαρακτηριστικά παραδείγματα είναι το `href` γνώρισμα ενός υπερσυνδέσμου (`<A>`) ή το `src` μιας εικόνας (`<IMG>`). Για τους παραπάνω λόγους, χρησιμοποιήθηκαν κατάλληλες

μεθοδολογίες ώστε το πρόγραμμα να έχει την ικανότητα εξαγωγής πολλαπλών στιγμιότυπων επιθυμητής πληροφορίας, με πολλαπλά πεδία το καθένα.

Δίνεται λοιπόν η δυνατότητα καθορισμού από το χρήστη του ρόλου που παίζει κάθε κόμβος του δενδρικού κανόνα εξαγωγής. Έτσι, ο χρήστης μπορεί να επιλέξει μεταξύ έξι διαφορετικών καταστάσεων λειτουργίας (states). Κάθε κατάσταση εκφράζει το κατά πόσο ο κόμβος πρέπει οπωσδήποτε να βρίσκεται σε ένα στιγμιότυπο επιθυμητής πληροφορίας, δηλαδή αν είναι *υποχρεωτικός* (required) ή *προαιρετικός* (optional), καθώς και το εάν ο χρήστης ενδιαφέρεται να εξάγει πληροφορία από τον συγκεκριμένο κόμβο. Αρκεί λοιπόν ο χρήστης να επιλέξει από το τοπικό μενού του κόμβου την κατάσταση που επιθυμεί για τον κόμβο αυτό (Εικόνα 10).



Εικόνα 10: Επιλογή κατάστασης κόμβου στο τοπικό μενού του κανόνα εξαγωγής

Ο δυνατές καταστάσεις είναι λοιπόν οι εξής:

- `checked`: ο κόμβος αυτός είναι *υποχρεωτικός* σε ένα στιγμιότυπο επιθυμητής πληροφορίας και ο χρήστης επιθυμεί πληροφορία που αυτός περιέχει. Αποτελεί δηλαδή μία μεταβλητή εξόδου. Αν πρόκειται για `TEXT` κόμβο η πληροφορία που εξάγεται είναι το κείμενο που εσωκλείει (*inner text*), για υπερσύνδεσμο (`<A>`) αυτό που ενδιαφέρει είναι το `href` γνώρισμα, για `FORM` και `INPUT` στοιχεία το `name` γνώρισμα τους ενώ για στοιχεία εικόνων εξάγεται η τιμή του `src` γνωρίσματος. Για τα *υπόλοιπα* HTML στοιχεία εξάγεται το *εσωτερικό* τους κείμενο. Επίσης, σε περίπτωση που για ένα κόμβο υπάρχει κανονική έκφραση, τότε εξάγεται η συμβολοσειρά που ταίριαξε με το πρότυπο. Σε περίπτωση που η κανονική έκφραση περιέχει παρενθέσεις, τότε εξάγεται το αλφαριθμητικό που προκύπτει από τη συγχώνευση των τμημάτων της συμβολοσειράς στόχου που συμφώνησε με κάθε έκφραση σε παρενθέσεις.

-  `checkedSource`: ο HTML κόμβος είναι *υποχρεωτικός* σε ένα στιγμιότυπο και ο χρήστης επιθυμεί τον *πηγαίο* κώδικα του στοιχείου (*outer HTML*). Ένας κόμβος τέτοιου τύπου συνιστά μεταβλητή εξόδου.
-  `grayed`: ο κόμβος αυτός είναι *υποχρεωτικός* σε ένα στιγμιότυπο αλλά ο χρήστης *δεν* επιθυμεί δεδομένα από αυτόν.
-  `unchecked`: *δεν* ενδιαφέρει καθόλου η παρουσία του σε ένα στιγμιότυπο. Θα μπορούσε και να διαγραφεί τελείως. Η ύπαρξη του στον κανόνα, εξυπηρετεί κυρίως τη δυνατότητα εύκολης μελλοντικής ενσωμάτωσης του, αν και όταν αυτό καταστεί αναγκαίο.
-  `grayed_implied`: ο κόμβος αυτός είναι *προαιρετικός* σε ένα στιγμιότυπο και ο χρήστης *δεν* επιθυμεί κάποια πληροφορία από τον συγκεκριμένο κόμβο. Κατά συνέπεια, σε περίπτωση που ο κόμβος αυτός έχει παιδιά, το υποδένδρο με ρίζα τον κόμβο αυτό είναι προαιρετικό, άσχετα με τις επιμέρους καταστάσεις των υπόλοιπων κόμβων που περιέχονται στο υποδένδρο αυτό.
-  `checked_implied`: ο κόμβος είναι μεν *προαιρετικός* σε ένα στιγμιότυπο *αλλά* αν τυχόν ταιριάξει με ένα κόμβο του myDOM, τότε γίνεται εξαγωγή της πληροφορίας που αυτός περιέχει, όπως ειπώθηκε και στην κατάσταση `checked`. Αποτελεί λοιπόν και αυτός μεταβλητή εξόδου. Όσο για το υποδένδρο του, ισχύει ό,τι και για ένα κόμβο τύπου `grayed_implied`.

Επίσης, ο χρήστης μπορεί να διαγράψει και εντελώς ένα κόμβο και κατά συνέπεια και το υποδένδρο του. Καθώς γίνονται αλλαγές στον κανόνα, δίνεται η ευχέρεια στο χρήστη να κρατάει αντίγραφα (snapshots) του τρέχοντος κανόνα, τα οποία αποθηκεύονται στην καρτέλα *Snapshots* στην περιοχή 5 της εφαρμογής (Εικόνα 1). Φυσικά, ο χρήστης έχει τη δυνατότητα να επαναφέρει ένα τέτοιο αντίγραφο και να το καταστήσει τρέχον πρότυπο μέσω επιλογής στο τοπικό αναδυόμενο μενού. Όταν ο χρήστης διαγράφει εντελώς ένα κόμβο, τότε δημιουργείται αυτόματα ένα αντίγραφο του κανόνα.

Όπως προαναφέρθηκε, τις περισσότερες φορές δεν είναι τόσο απλό να εντοπιστούν όλα τα στιγμιότυπα και συνήθως χρειάζεται επεξεργασία του κανόνα και προσεκτική επιλογή των κατάλληλων καταστάσεων. Ιδιαίτερα χρήσιμη είναι μια

δευτερεύουσα βοηθητική (auxiliary) δενδρική δομή, στην οποία μπορεί ο χρήστης να τοποθετήσει ένα στιγμιότυπο που «έχασε» ο κανόνας. Αυτό γίνεται με διαδικασία παρόμοια με τη δημιουργία του κανόνα, απλά αυτή τη φορά ο χρήστης πρέπει να επιλέξει τη λειτουργία κατασκευής βοηθητικού στιγμιότυπου. Η δομή αυτή βρίσκεται ακριβώς αριστερά από τον τρέχοντα κανόνα. Έτσι, ο χρήστης έχοντας τις δύο αυτές δενδρικές δομές, τη μία δίπλα στην άλλη, μπορεί εύκολα να εντοπίσει τις διαφορές τους και συνεπώς το λόγο για τον οποίο ο κανόνας δεν εξήγαγε κάποιο στιγμιότυπο. Συμπερασματικά, η δομή αυτή διευκολύνει αρκετά την κατασκευή και την κατάλληλη ρύθμιση αποτελεσματικών κανόνων εξαγωγής.

Πλοήγηση στην Επόμενη Σελίδα

Μια τυπική σελίδα αποτελεσμάτων μιας μηχανής αναζήτησης ή μιας υπηρεσίας σύγκρισης τιμών περιέχει πολλά στιγμιότυπα επιθυμητής πληροφορίας. Πολύ συχνά μάλιστα το πλήθος των αποτελεσμάτων είναι μεγάλο και συνεπώς αυτά βρίσκονται μοιρασμένα σε πολλές διαδοχικές ιστοσελίδες, οι οποίες διασυνδέονται μεταξύ τους μέσω υπερσυνδέσμων με κείμενο τύπου '`Next`'. Οι wrappers που φτιάχνονται με το ΔEiXTo έχουν τη δυνατότητα, με χρήση ενός απλού μηχανισμού, να εντοπίζουν τον υπερσύνδεσμο που οδηγεί στην εκάστοτε επόμενη σελίδα και εφόσον αυτός βρεθεί, τον ακολουθούν και συνεχίζουν τη συλλογή των αποτελεσμάτων στη νέα σελίδα.

Ο χρήστης μπορεί να ενεργοποιήσει τη λειτουργία αυτή μέσω στοιχείων ελέγχου της καρτέλας *Project Info* στην περιοχή 5 της εφαρμογής (Εικόνα 1). Τα στοιχεία αυτά απεικονίζονται στην Εικόνα 11.

Εικόνα 11: Στοιχεία ελέγχου για πλοήγηση σε επόμενη σελίδα

Το ΔEiXTo επιτρέπει τον εντοπισμό του συνδέσμου που οδηγεί στην επόμενη σελίδα είτε με χρήση του εσωκλειόμενου κειμένου του είτε με βάση τον τίτλο του. Απαραίτητη προϋπόθεση προφανώς για την καλή απόδοση του μηχανισμού αυτού είναι η ύπαρξη κειμένου ή γνωρίσματος title στον ζητούμενο υπερσύνδεσμο, γεγονός που στη συντριπτική πλειοψηφία των περιπτώσεων τέτοιων σελίδων ισχύει. Ο χρήστης έχει τη δυνατότητα να εισάγει μέρος (πρόθεμα) του κειμένου ή του τίτλου που έχει ο '`Next`' σύνδεσμος, ώστε να γίνει δυνατός ο προσδιορισμός του μεταξύ των

υπόλοιπων υπερσυνδέσμων της σελίδας. Για ευκολία, η σύγκριση που γίνεται μεταξύ της συμβολοσειράς που δίνει ο χρήστης και του εσωτερικού κειμένου ή του τίτλου κάθε συνδέσμου, για τον εντοπισμό του συνδέσμου που δείχνει στην επόμενη σελίδα, δε διαχωρίζει πεζά από κεφαλαία γράμματα (case insensitive). Επίσης, ο χρήστης μπορεί να ορίσει και το βάθος πλοήγησης (crawling depth), δηλαδή το μέγιστο πλήθος των σελίδων που θα επισκεφτεί ο wrapper ακολουθώντας '`Next`' συνδέσμους.

Χρήση Κανονικών Εκφράσεων

Πολλές φορές είναι χρήσιμο να θέσουμε *περιορισμούς* για το περιεχόμενο ορισμένων κόμβων του κανόνα, ώστε να διευκολυνθεί ο εντοπισμός των σωστών στιγμιότυπων επιθυμητής πληροφορίας. Για παράδειγμα, μπορεί ο χρήστης να θέλει να ορίσει ότι για να ταιριάξει ένας κόμβος κειμένου του myDOM με κάποιον του προτύπου, θα πρέπει να ξεκινάει με ένα συγκεκριμένο πρόθεμα ή να περιέχει μια συγκεκριμένη συμβολοσειρά. Άλλες φορές πάλι ο χρήστης μπορεί να θέλει να απομονώσει ένα ή και περισσότερα τμήματα του κειμένου που περιέχει ένας κόμβος. Τα παραπάνω επιτυγχάνονται με χρήση κανονικών εκφράσεων (regular expressions), οι οποίες συνιστούν μία τυπική μέθοδο περιγραφής προτύπων κειμένου. Μία σύντομη αλλά περιεκτική παρουσίαση των κανονικών εκφράσεων γίνεται στο Παράρτημα. Για να γίνει καλύτερα κατανοητή η σημασία της χρήσης τους στο ΔEiXTo ακολουθούν δύο απλά παραδείγματα.

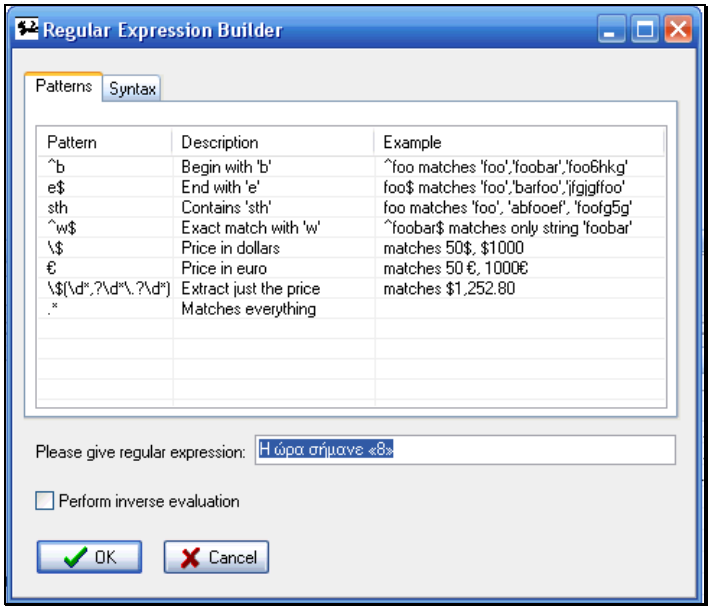
Έστω λοιπόν ότι ένας χρήστης επιθυμεί να εξάγει τη διεύθυνση (`URL`) στην οποία οδηγεί ένας υπερσύνδεσμος με κείμενο '`Next`'. Δεδομένου ότι οι περισσότεροι σύνδεσμοι μιας σελίδας έχουν τη δομή που φαίνεται στην Εικόνα 12, δεν αρκεί ο κανόνας που προκύπτει απλά με την οπτική υπόδειξη του επιθυμητού συνδέσμου. Με χρήση του κανόνα της ακόλουθης εικόνας επιστρέφονται τα `href` γνωρίσματα όσων συνδέσμων έχουν αυτή τη δομή, δηλαδή σχεδόν όλων. Προφανώς, αυτό δεν είναι το επιθυμητό αποτέλεσμα της εξαγωγής.

Εικόνα 12: Τυπική δομή υπερσυνδέσμου

Αν όμως ο χρήστης ορίσει για τον `TEXT` κόμβο κανονική έκφραση της μορφής '`Next`', τότε ο κανόνας επιστρέφει μόνο τον επιθυμητό υπερσύνδεσμο, καθώς το πρότυπο θα συμφωνήσει μόνο με τον ζητούμενο υπερσύνδεσμο.

Ένα άλλο παράδειγμα που αναδεικνύει τη χρησιμότητα των κανονικών εκφράσεων προέρχεται από περιπτώσεις όπου ο χρήστης θέλει να απομονώσει κάποια συγκεκριμένη πληροφορία που περιέχεται μέσα στο κείμενο ενός `TEXT` κόμβου. Έστω λοιπόν ότι ένας συγκεκριμένος `TEXT` κόμβος των στιγμιότυπων της επιθυμητής πληροφορίας περιέχει κείμενο της μορφής '`from $249.98`' και ο χρήστης ενδιαφέρεται μόνο για την ακέραια αριθμητική τιμή. Τότε, δεν έχει παρά να αναθέσει στον αντίστοιχο κόμβο του κανόνα, μία κανονική έκφραση όπως αυτή, `\$(\d+)`.

Να σημειωθεί πως δυνατότητα εφαρμογής κανονικής έκφρασης υπάρχει για *όλους* τους τύπους κόμβων του κανόνα, δηλαδή τόσο για τους `TEXT` κόμβους όσο και για τους HTML. Η ανάθεση κανονικής έκφρασης σε ένα κόμβο του κανόνα γίνεται με κατάλληλη επιλογή από το τοπικό μενού του κόμβου. Ο χρήστης μπορεί να επιλέξει στο σχετικό παράθυρο (Εικόνα 13) από προκατασκευασμένες κανονικές εκφράσεις που αφορούν συνήθεις περιπτώσεις ή να γράψει μία νέα. Για την απομόνωση ενός ή περισσότερων τμημάτων του περιεχομένου ενός κόμβου, απαιτείται χρήση παρενθέσεων στην κανονική έκφραση. Επίσης, υπάρχει δυνατότητα για αντίστροφη εφαρμογή (inverse evaluation) της κανονικής έκφρασης, δηλαδή χρήση της άρνησης της δοθείσας κανονικής έκφρασης. Αρκεί να ενεργοποιηθεί η σχετική λειτουργία στο ίδιο παράθυρο. Οι κόμβοι που φέρουν κανονική έκφραση φαίνονται με υπογραμμισμένα γράμματα. Για απομάκρυνση της κανονικής έκφρασης και επαναφορά ενός κόμβου στην αρχική του κατάσταση, πρέπει να επιλεγεί από το τοπικό του μενού η αντίστοιχη λειτουργία.



Εικόνα 13: Παράθυρο διαλόγου για ανάθεση κανονικής έκφρασης

Αξίζει, τέλος, να τονιστεί ότι οι κανονικές εκφράσεις παρέχουν μερική δυνατότητα για διατύπωση *μαθηματικών περιορισμών*. Για παράδειγμα, η έκφραση `[7-9]\d\d`, εκφράζει όλους τους αριθμούς από `700` έως και `999`. Η λειτουργία αυτή μπορεί να φανεί αρκετά χρήσιμη, ιδιαίτερα σε σελίδες με καταναλωτικά αγαθά, οι οποίες μάλιστα συγκεντρώνουν και μεγάλο ενδιαφέρον από τους χρήστες του Παγκόσμιου Ιστού.

Εκτέλεση Κανόνα Εξαγωγής

Οι κανόνες εξαγωγής απλά περιγράφουν τη μορφή της πληροφορίας που θα εξαχθεί και όχι τον τρόπο με τον οποίο αυτό θα γίνει. Άρα είναι απαραίτητη η ύπαρξη ενός *μηχανισμού εκτέλεσης* (executor), ο οποίος και υλοποιεί την εξαγωγή δεδομένων με βάση το πρότυπο που έχει ορίσει ο χρήστης. Στο ΔΕiΧΤο, για να εκτελεστεί ένας κανόνας στην προβαλλόμενη ιστοσελίδα, πρέπει ο χρήστης να πατήσει το σχετικό κουμπί (Εικόνα 14) στην περιοχή 4 της εφαρμογής (Εικόνα 1).

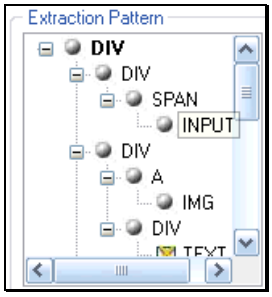
Εικόνα 14: Κουμπί εκτέλεσης κανόνα εξαγωγής

Σε ορισμένες περιπτώσεις, συνήθως όταν η εκτέλεση του κανόνα περιλαμβάνει πλοήγηση σε πολλαπλές '`Next`' σελίδες, χρήσιμη είναι η δυνατότητα διακοπής της εκτέλεσης από το χρήστη με το πάτημα σχετικού κουμπιού (Εικόνα 15) στην περιοχή 2 της εφαρμογής.

Εικόνα 15: Κουμπί διακοπής εκτέλεσης κανόνα εξαγωγής

Όταν δοθεί εντολή εκτέλεσης από το χρήστη, τότε δημιουργείται ένα αντίγραφο του κανόνα, χωρίς τους `unchecked` κόμβους, το οποίο φαίνεται στη δενδρική δομή της καρτέλας *Project Info* (Εικόνα 16). Το αντίγραφο αυτό, του οποίου οι κόμβοι δεν περιέχουν δεδομένα, συνιστά το πρότυπο αναζήτησης επιθυμητής πληροφορίας. Συγκεκριμένα, γίνεται προσπάθεια συμφωνίας του προτύπου στη myDOM αναπαράσταση, δηλαδή γίνεται αναζήτηση υποδένδρων που ταιριάζουν με το πρότυπο, ώστε να εντοπιστούν τα επιθυμητά δεδομένα μέσα στο

περιεχόμενο της υπό εξέταση σελίδας. Για το σκοπό αυτό ελέγχονται όλοι οι κόμβοι της myDOM αναπαράστασης της σελίδας.



Εικόνα 16: Δενδρική δομή για πρότυπο εξαγωγής

Αν είναι ενεργοποιημένη η λειτουργία πλοήγησης σε 'Next' σελίδες, τότε η ίδια διαδικασία συνεχίζεται και στις επόμενες σελίδες, ώστε να συγκεντρωθούν όλα τα επιθυμητά στιγμιότυπα. Ο αλγόριθμος συμφωνίας προτύπου περιγράφεται αναλυτικά σε επόμενη ενότητα.

Όταν βρεθεί ένα ταίριασμα με το πρότυπο, δηλαδή όλοι οι περιορισμοί που ορίζει ο κανόνας ικανοποιηθούν για ένα υποδένδρο του myDOM, τότε μέρος του περιεχομένου της ιστοσελίδας θα έχει δεσμευτεί στις παραμέτρους εξόδου του κανόνα, δηλαδή στους κόμβους του κανόνα που είναι σε κατάσταση `checked` ή `checkedSource` ή `checked_implied`. Τα προκαθορισμένα ονόματα των μεταβλητών είναι τύπου `VARX` (π.χ. *VAR1*, *VAR2*, *VAR3*, κ.ο.κ.). Στο σημείο αυτό, ο μηχανισμός εκτέλεσης συλλέγει τις τιμές από τις μεταβλητές εξόδου και δημιουργεί μια εγγραφή εξόδου (output record) ή αλλιώς ένα αποτέλεσμα εξαγωγής (extraction result). Σε περίπτωση που το υποδέντρο του myDOM με ρίζα το υπό εξέταση στοιχείο δεν ταιριάξει με το πρότυπο, τότε απορρίπτονται τα περιεχόμενα του τρέχοντος αντίγραφου του κανόνα και αρχίζει ένα νέος κύκλος ώστε να ελεγχθεί το επόμενο στοιχείο. Η αναζήτηση ταιριασμάτων στην προβαλλόμενη σελίδα τερματίζει όταν ελεγχθούν όλα τα στοιχεία της myDOM αναπαράστασής της.

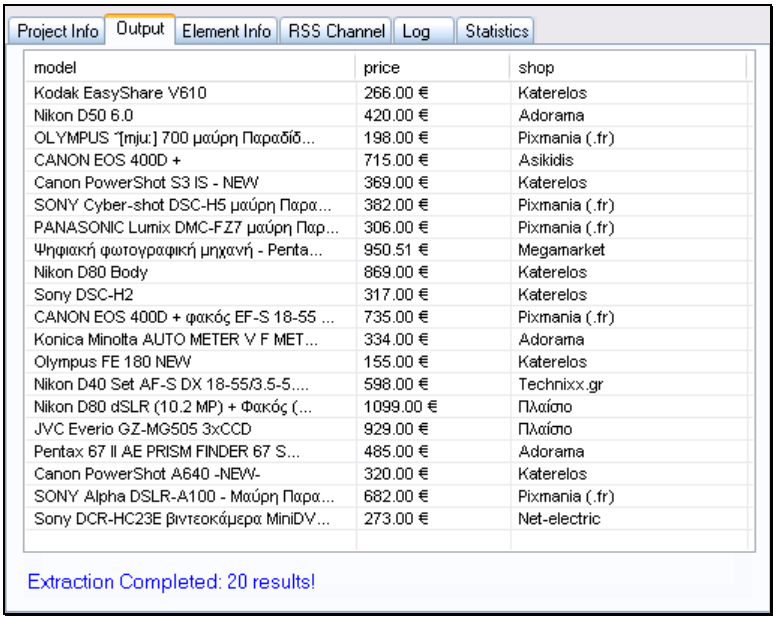
Τα αποτελέσματα της εκτέλεσης ενός κανόνα, τυπώνονται σε πίνακα στην καρτέλα *Output*, η οποία βρίσκεται στην περιοχή 5 της εφαρμογής Οι στήλες του πίνακα είναι τόσες όσες και οι μεταβλητές εξόδου και ο αριθμός των γραμμών προφανώς ισοδυναμεί με το πλήθος των συμφωνιών προτύπου, δηλαδή των στιγμιότυπων που εντοπίστηκαν. Τα ονόματα των στηλών είναι ίδια με τα προκαθορισμένα αναγνωριστικά των μεταβλητών εξόδου, δηλαδή `VARX`. Ωστόσο, ο χρήστης μπορεί να αλλάξει την ετικέτα (label) μιας μεταβλητής εξόδου μέσω

σχετικής επιλογής στο τοπικό μενού του επιθυμητού κόμβου του κανόνα εξαγωγής, ώστε στον πίνακα αποτελεσμάτων η αντίστοιχη στήλη να αποκτήσει όνομα σχετικό με την εξαγόμενη πληροφορία.

Η ετικέτα που εισάγει ο χρήστης, μέσω παράθυρου διαλόγου, συνδυάζεται με το όνομα του κόμβου και σαν διαχωριστικός χαρακτήρας χρησιμοποιείται το `':'`, όπως φαίνεται στην Εικόνα 17. Έτσι, αν και χειροκίνητα, γίνεται εφικτή η περιγραφή της σημασιολογίας του χρήσιμου περιεχομένου μιας σελίδας. Στην Εικόνα 18 φαίνονται κάποια ενδεικτικά αποτελέσματα εξαγωγής δεδομένων από μία ιστοσελίδα υπηρεσίας σύγκρισης τιμών.

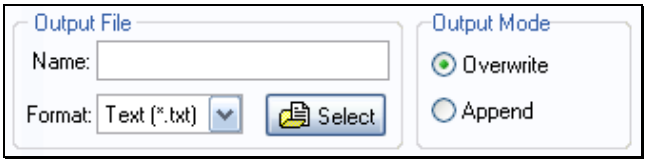
Εικόνα 17: Κόμβος κανόνα με ετικέτα

Επίσης, η επιλογή μίας εγγραφής του πίνακα αποτελεσμάτων χρωματίζει διαφορετικά στην προβαλλόμενη σελίδα, εφόσον αυτό είναι δυνατό, το στιγμιότυπο στο οποίο αντιστοιχεί το συγκεκριμένο αποτέλεσμα. Αυτό είναι αρκετά χρήσιμο για τον εντοπισμό των στιγμιότυπων που ο κανόνας πιθανώς δεν “έπιασε”. Ακόμα, με διπλό κλικ σε ένα αποτέλεσμα, ανοίγει η ιστοσελίδα στην οποία βρίσκεται το αντίστοιχο στιγμιότυπο, σε νέο παράθυρο του Internet Explorer. Αυτό έχει νόημα στις περιπτώσεις όπου ο wrapper εφαρμόζει τον κανόνα εξαγωγής σε πολλαπλές σελίδες, οπότε τα αποτελέσματα που συγκεντρώνει, προέρχονται από πολλές διαφορετικές διευθύνσεις. Έτσι, διευκολύνεται και η επαλήθευση των αποτελεσμάτων.



Εικόνα 18: Αποτελέσματα από εκτέλεση ενδεικτικού κανόνα

Τα αποτελέσματα από μία εκτέλεση μπορούν να εξαχθούν σε αρχείο (export data). Υποστηρίζονται τρεις διαφορετικοί τύποι αρχείων εξόδου, *απλό κείμενο* (*tab delimited text*), *XML* και *RSS*. Για τους δύο τελευταίους τύπους γίνεται χρήση των ονομάτων των μεταβλητών εξόδου και συνεπώς των ετικετών που έχει εισάγει ο χρήστης. Έτσι, κάθε ετικέτα που έχει δώσει ο χρήστης, χρησιμοποιείται ως XML τύπος στοιχείου. Για την εξαγωγή των αποτελεσμάτων του wrapper σε αρχείο, πρέπει ο χρήστης να πάει στην καρτέλα *Project Info*, να επιλέξει τύπο αρχείου, να δώσει το όνομα ή/και τη διαδρομή (path) του αρχείου εξόδου και τον τρόπο αποθήκευσης (mode), δηλαδή απόρριψη τυχόν προϋπαρχόντων περιεχομένων (overwrite) του αρχείου ή προσθήκη στο τέλος του αρχείου (append). Τα σχετικά στοιχεία ελέγχου της εφαρμογής φαίνονται στην Εικόνα 19.



Εικόνα 19: Στοιχεία ελέγχου για έξοδο σε αρχείο

Να σημειωθεί πως λαμβάνεται μέριμνα ώστε τα αρχεία εξόδου να έχουν πάντα κωδικοποίηση (encoding) χαρακτήρων *UTF-8*, ανεξάρτητα από την κωδικοποίηση των ιστοσελίδων. Έτσι, αν ο χρήστης επέλεγε για παράδειγμα τα αποτελέσματα που απεικονίζονται στην Εικόνα 18 να εξαχθούν σε XML αρχείο, τότε θα προέκυπτε αρχείο της μορφής που φαίνεται στην Εικόνα 20.

```
<?xml version="1.0" encoding="UTF-8" ?>
- <WrapperResults>
  - <item>
      <model>Kodak EasyShare V610</model>
      <price>266.00 €</price>
      <shop>Katerelos</shop>
    </item>
  - <item>
      <model>Nikon D50 6.0</model>
      <price>420.00 €</price>
      <shop>Adorama</shop>
    </item>
```

Εικόνα 20: Τμήμα ενδεικτικού XML αρχείου εξόδου

Στις περιπτώσεις εξαγωγής δεδομένων σε tab delimited ή XML αρχείο, υπάρχει η δυνατότητα συμπερίληψης σε κάθε εγγραφή που εξάγεται, της διεύθυνσης της σελίδας από την οποία αυτή προήλθε (native URL). Αρκεί να ενεργοποιηθεί η σχετική λειτουργία στην καρτέλα *Project Info*. Επιπλέον, είναι δυνατός ο ορισμός

μέγιστου πλήθους αποτελεσμάτων, το οποίο μπορεί να λειτουργήσει ως κριτήριο τερματισμού εκτέλεσης ενός κανόνα. Οι δύο αυτές λειτουργίες αποτυπώνονται στην Εικόνα 21.

Εικόνα 21: Μέγιστο πλήθος αποτελεσμάτων και συμπερίληψη native URL

Πρέπει επίσης να σημειωθεί πως κατά την εκτέλεση ενός κανόνα, το πρόγραμμα βρίσκεται σε *κατάσταση εκτέλεσης* (running mode) και όλα τα γεγονότα (events) του προγράμματος πλοήγησης που προκαλούνται από το χρήστη μέσω mouse, απενεργοποιούνται έως ότου η εκτέλεση να διεκπεραιωθεί. Για παράδειγμα, κατά τη διάρκεια της εκτέλεσης, ο χρήστης δεν μπορεί να ακολουθήσει ένα υπερσύνδεσμο μέσω αριστερού κλικ, ούτε μπορεί να εμφανίσει αναδυόμενο μενού με δεξί κλικ. Αυτό είναι κυρίως χρήσιμο στην εκτέλεση wrapper που περιλαμβάνει πλοήγηση σε πολλαπλές '`Next`' σελίδες και γίνεται για να εξασφαλιστεί η ομαλή εξαγωγή των επιθυμητών δεδομένων. Σε περίπτωση που το πρόγραμμα δεν βρίσκεται σε φάση εκτέλεσης, ο χρήστης μπορεί να χρησιμοποιεί κανονικά το ενσωματωμένο πρόγραμμα πλοήγησης.

Εικονική Ρίζα Κανόνα

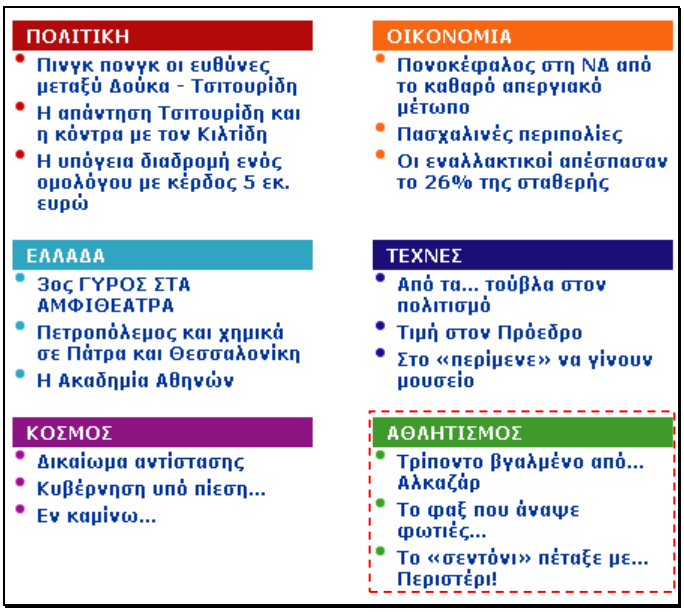
Σε ορισμένες περιπτώσεις η δομή του προτύπου μπορεί να είναι σχετικά απλή, με συνέπεια να επιστρέφονται λανθασμένα υπεράριθμα αποτελέσματα, καθώς η συγκεκριμένη δομή είναι πολύ διαδεδομένη στη σελίδα. Αυτό σημαίνει πως το πρότυπο πρέπει να γίνει πιο αυστηρό, δηλαδή να τεθούν ορισμένοι περιορισμοί. Οι κανονικές εκφράσεις κάποιες φορές βοηθούν αρκετά αλλά δεν αποτελούν πανάκεια. Συνήθως, για την επίλυση του προβλήματος, απαιτείται να περιγραφεί το περιβάλλον ή αλλιώς η γειτονιά του ριζικού κόμβου του κανόνα.

Στο ΔEiXTo αυτό επιτυγχάνεται με εισαγωγή στον κανόνα ορισμένων άμεσων, στο myDOM δέντρο, προγόνων (πατέρα, πατέρα του πατέρα, κ.ο.κ.) του τρέχοντος ριζικού κόμβου του κανόνα και ενδεχομένως προσθήκη αδελφικών κόμβων (siblings) τους. Συγκεκριμένα, ο χρήστης μπορεί να «ανεβοκατεβαίνει» επίπεδα μέσω των σχετικών κουμπιών (Εικόνα 22) που βρίσκονται στην περιοχή 4

του παραθύρου της εφαρμογής και να προσθέτει αδέλφια (siblings) σε ένα κόμβο πρόγονο της αρχικής ρίζας, με σχετική επιλογή από το τοπικό μενού του κόμβου. Να σημειωθεί πως η επιλογή εισαγωγής αδελφικού κόμβου εισάγει στον κανόνα το υποδένδρο με ρίζα τον κόμβο αυτό.

Εικόνα 22: Κουμπιά προσθήκης και αφαίρεσης επιπέδων κανόνα

Για παράδειγμα, σε ένα ενδεικτικό ειδησεογραφικό ιστοχώρο (Εικόνα 23), οι επικεφαλίδες των ειδήσεων έχουν τον ίδιο ακριβώς τρόπο εμφάνισης και είναι οργανωμένες ανά κατηγορία σε πίνακες. Έστω λοιπόν ότι ο χρήστης επιθυμεί τις αθλητικές ειδήσεις.



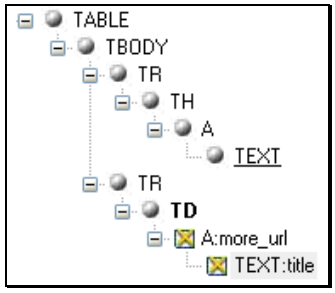
Εικόνα 23: Επικεφαλίδες ειδήσεων από ενδεικτικό ειδησεογραφικό ιστοχώρο

Προφανώς, όλες οι επικεφαλίδες έχουν κοινή HTML δομή η οποία φαίνεται στην Εικόνα 24, γεγονός που προκαλεί πρόβλημα στην απομόνωση και εξαγωγή μόνο της επιθυμητής πληροφορίας.

Εικόνα 24: Δομή επικεφαλίδας ειδήσεων

Για την επίλυση του προβλήματος απαιτείται να διευκρινιστεί το περιβάλλον της ρίζας του κανόνα εξαγωγής. Με χρήση των σχετικών λειτουργιών του εργαλείου,

μπορεί εύκολα να προκύψει ο κανόνας της Εικόνα 25, ο οποίος επιστρέφει μόνο τα επιθυμητά δεδομένα. Αξίζει να τονιστεί ότι ο κανόνας χρησιμοποιεί σαν *σημείο αναφοράς* (landmark) το όνομα της κατηγορίας ειδήσεων, 'ΑΘΛΗΤΙΣΜΟΣ' στην προκειμένη περίπτωση, το οποίο περιγράφεται με χρήση κανονικής έκφρασης.



Εικόνα 25: Κανόνας για επικεφαλίδες αθλητικών ειδήσεων

Όπως φαίνεται στην Εικόνα 25, ρίζα του κανόνα εξαγωγής είναι ο κόμβος `TABLE`, ενώ ρίζα του αρχικού κανόνα ήταν το `TD`, το οποίο φαίνεται σε έντονη γραφή. Το υποδένδρο με ρίζα το `TD` αναπαριστά ένα στιγμιότυπο επιθυμητής πληροφορίας και συνεπώς αυτό είναι το πρότυπο που πρέπει να αναζητηθεί μέσα στη δενδροειδή myDOM αναπαράσταση. Οι κόμβοι πάνω από το `TD` αποτελούν περιορισμούς γειτονιάς. Για το λόγο αυτό, κατά την εκτέλεση του παραπάνω κανόνα, αναζητούνται ταιριάσματα (matches) στο myDOM με πρότυπο το υποδένδρο του `TD` και έπειτα για κάθε ταίριασμα ελέγχεται η γειτονιά του. Μόνο αν πληρούνται όλοι οι περιορισμοί που εκφράζει ο κανόνας, υπάρχει επιτυχής συμφωνία με το πρότυπο.

Η τεχνική αυτή καλείται μέθοδος *εικονικής ρίζας* (virtual root), καθώς ως ρίζα της δομής που αναζητείται δε χρησιμοποιείται η πραγματική ρίζα του κανόνα αλλά ο ριζικός κόμβος του υποδένδρου που αντιστοιχεί στο στιγμιότυπο της επιθυμητής πληροφορίας.

Διαδοχικοί Προαιρετικοί Κόμβοι

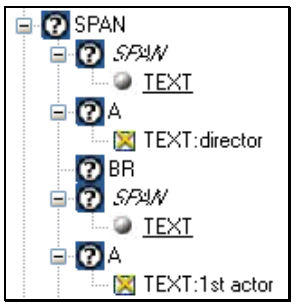
Κατά την προσπάθεια συμφωνίας προτύπου, ο αλγόριθμος όταν δεν ταιριάζει ένα προαιρετικό κόμβο του κανόνα, συνεχίζει με τον επόμενο αδελφικό του κόμβο, εφόσον βέβαια αυτός υπάρχει. Ωστόσο, σε ορισμένες περιπτώσεις, οι διαδοχικοί προαιρετικοί κόμβοι αποτελούν μία ομάδα (group), που σημαίνει ότι «πηγαίνουν» πάντα μαζί. Έτσι, μερικές φορές είναι χρήσιμο ο αλγόριθμος να τους χειρίζεται σαν ομάδα κόμβων. Δηλαδή, σε περίπτωση που δεν ταιριάξει ένα προαιρετικό κόμβο, να προσπεράσει καθορισμένο αριθμό επόμενων αδελφικών κόμβων. Αυτό γίνεται

δυνατό με την ύπαρξη ενός αριθμού *FSON* (Following Successive Optional Nodes) σε ένα προαιρετικό κόμβο που εκφράζει το πλήθος των επόμενων του διαδοχικών προαιρετικών κόμβων. Ο χρήστης μπορεί να δώσει τιμή σε αυτόν μέσω της σχετικής επιλογής στο τοπικό μενού του κόμβου.

Για παράδειγμα, θεωρούμε την περίπτωση της εγγραφής στην Εικόνα 26 και κάνουμε την παραδοχή ότι τα τμήματα που αφορούν στο σκηνοθέτη και τους ηθοποιούς είναι προαιρετικά, σε αντίθεση με τον τίτλο της ταινίας που είναι υποχρεωτικό να υπάρχει σε ένα στιγμιότυπο.

Εικόνα 26: Εγγραφή με προαιρετικά τμήματα δεδομένων

Στον κανόνα που θα κατασκεύαζε ο χρήστης, θα έπρεπε τους κόμβους που αντιστοιχούν στο τμήμα δεδομένων '`Director: όνομα_σκηνοθέτη`' και στο '`Actors: ονόματα_ηθοποιών`' να τους θέσει ως προαιρετικούς, όπως απεικονίζεται στην Εικόνα 27. Ο πρώτος `TEXT` κόμβος περιέχει την κανονική έκφραση `Director` ενώ ο δεύτερος την έκφραση `Actors`.



Εικόνα 27: Τμήμα κανόνα με διαδοχικούς προαιρετικούς κόμβους

Όμως, σε στιγμιότυπο που *δεν* δίνεται ο σκηνοθέτης, το ταίριασμα του πρώτου `SPAN` θα αποτύγχανε αλλά ο κόμβος `A` του σκηνοθέτη θα ταίριαζε με τον myDOM κόμβο `A` του πρώτου ηθοποιού του στιγμιότυπου. Το γεγονός αυτό συνιστά σφάλμα καθώς το συγκεκριμένο `A` του κανόνα πηγαίνει πάντα μαζί με τους αδελφικούς του κόμβους, `SPAN` και `BR`, που αφορούν στο σκηνοθέτη.

Για την επίλυση του παραπάνω προβλήματος παρέχεται η δυνατότητα ελεγχόμενου χειρισμού των διαδοχικών προαιρετικών κόμβων σαν ομάδα. Έτσι, στην προκειμένη περίπτωση, αν δοθεί τιμή 2 στο πεδίο `FSON` του `SPAN` του σκηνοθέτη, τότε

αν αυτό δε βρεθεί, ο αλγόριθμος μεταπηδά στο SPAN των ηθοποιών καθώς προσπερνά τους επόμενους δύο αδελφικούς κόμβους (A και BR).

Αλγόριθμος Συμφωνίας Προτύπου

Ο αλγόριθμος που χρησιμοποιεί το ΔΕiΧΤο για συμφωνία (ή ταίριασμα) προτύπου (pattern matching) αποδεικνύεται ιδιαίτερα αποτελεσματικός, τουλάχιστον για την πλειοψηφία των περιπτώσεων. Όπως περιγράφηκε και σε προηγούμενη ενότητα, ένας κανόνας εξαγωγής μπορεί να έχει εικονική ρίζα, με συνέπεια το δέντρο του να «σπάει» σε δύο τμήματα. Τις περισσότερες φορές βέβαια, η εικονική και η αληθινή ρίζα του προτύπου συμπίπτουν. Για καλύτερη περιγραφή του αλγορίθμου γίνονται ορισμένες συμβάσεις ονοματολογίας. Έστω R το δέντρο του κανόνα και vroot ο κόμβος της εικονικής ρίζας του R. Τότε, T1 ονομάζεται το υποδένδρο του κανόνα με ρίζα τον κόμβο vroot, ενώ T2 καλείται το υποδένδρο R-T1, το οποίο αποτελείται από τους κόμβους που βρίσκονται «πάνω» από τον κόμβο vroot. Το T2 συνιστά τη γειτονιά του vroot. Σε περίπτωση που η εικονική ρίζα ταυτίζεται με την πραγματική ρίζα του δέντρου του κανόνα, τότε απλά θεωρείται ότι το T2 είναι κενό.

Για τον εντοπισμό στιγμιότυπων επιθυμητής πληροφορίας, γίνεται προσπέλαση κάθε στοιχείου της myDOM αναπαράστασης της προβαλλόμενης σελίδας. Σε κάθε myDOM κόμβο αντιστοιχεί ένα νέος κύκλος προσπάθειας συμφωνίας με το πρότυπο. Έστω node ο κόμβος του myDOM υπό εξέταση και S το υποδένδρο με ρίζα τον κόμβο αυτό. Ο αλγόριθμος συμφωνίας προτύπου αποτελείται από δύο βασικά στάδια. Στο πρώτο στάδιο ελέγχεται αν το S ταιριάζει με το T1 και στο δεύτερο ελέγχεται αν η γειτονιά του node ταιριάζει με το T2. Αν και οι δύο έλεγχοι είναι επιτυχείς, δηλαδή ικανοποιηθούν όλοι οι περιορισμοί που επιβάλλει το πρότυπο, τότε υπάρχει ταίριασμα, που σημαίνει ότι εντοπίστηκε στιγμιότυπο επιθυμητής πληροφορίας, οπότε και εξάγονται τα δεδομένα του.

Η κεντρική ιδέα πίσω από τον αλγόριθμο που χρησιμοποιεί το ΔΕiΧΤο είναι ότι για να ταιριάξουν δύο κόμβοι πρέπει να έχουν την ίδια ετικέτα και να ταιριάζουν οι θυγατρικοί τους κόμβοι. Το ταίριασμα δύο δέντρων συνεπώς ανάγεται σε πρόβλημα ταιριάσματος των ριζικών τους κόμβων και προφανώς επιτυγχάνεται μέσω αναδρομής (recursion) και σάρωσης δέντρου κατά βάθος. Βασικά χαρακτηριστικά του αλγόριθμου είναι η υποστήριξη ενδεχόμενης απουσίας κόμβων (missing nodes) στο δέντρο στόχο και η ύπαρξη προαιρετικών τμημάτων στον κανόνα. Η διαδικασία

ταιριάσματος ενός κόμβου του S με ένα κόμβο P του T1 βασίζεται στην αρχή της *πρώτης εμφάνισης* (first occurrence). Δηλαδή, σε ένα κύκλο ο αλγόριθμος διασχίζει τους κόμβους του αντίστοιχου επιπέδου του S και σταματά την αναζήτηση ταιριάσματος για τον P, όταν βρει τον πρώτο κόμβο του S που συμφωνεί με τον P. Να σημειωθεί πως η προσπάθεια ταιριάσματος ενός κόμβου γίνεται από το σημείο που συνέβη το τελευταίο ταίριασμα.

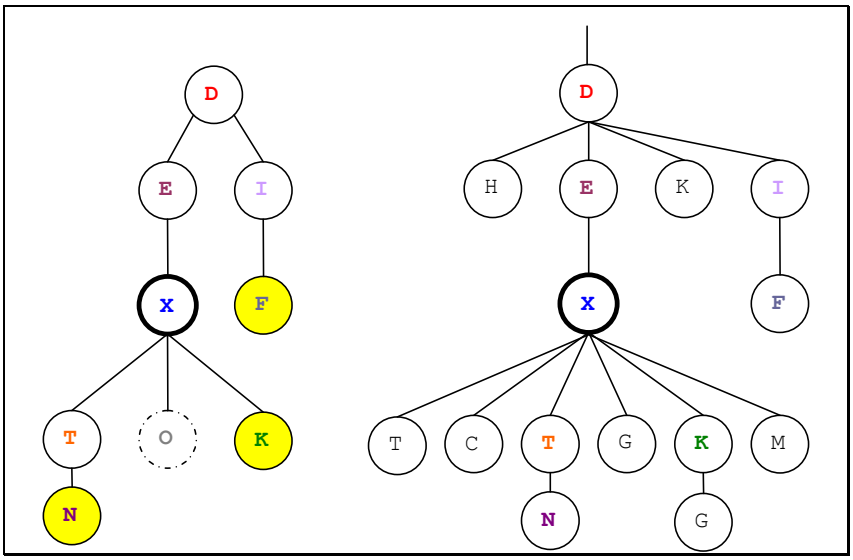
Συνεπώς, όταν ταιριάζει ένας κόμβος του myDOM με ένα κόμβο του προτύπου, σημαίνει ότι έχουν ταιριάξει προηγουμένως και τα υποδένδρα τους. Μάλιστα, όταν γίνεται συμφωνία ενός κόμβου, τότε το περιεχόμενο του αντίστοιχου κόμβου του προτύπου γεμίζει με τα δεδομένα του κόμβου με τον οποίο ταίριαξε. Έτσι, σε περίπτωση που βρεθεί συνολικό ταίριασμα με το πρότυπο, τότε όλοι οι κόμβοι του προτύπου έχουν αποκτήσει δεδομένα, ορισμένα εκ των οποίων είναι αυτά που επιθυμεί ο χρήστης, οπότε και εξάγονται.

Σε περίπτωση που δεν ταιριάξει ένας υποχρεωτικός κόμβος του προτύπου, τότε η διαδικασία αποτυγχάνει και έτσι ξεκινά ένας νέος κύκλος προσπάθειας συμφωνίας προτύπου με τον επόμενο myDOM κόμβο. Αν δε βρεθεί ταίριασμα για ένα προαιρετικό κόμβο του προτύπου, τότε το υποδένδρο του μένει κενό από δεδομένα και ο αλγόριθμος συνεχίζει κανονικά με τον επόμενο αδελφικό του κόμβο, εφόσον αυτός υπάρχει. Μάλιστα, η αναζήτηση ταιριάσματος στο δέντρο στόχο για το νέο κόμβο συνεχίζει από τον κόμβο που ταίριαξε τελευταίος. Σε περίπτωση βέβαια που είναι ενεργοποιημένος ο χειρισμός των διαδοχικών προαιρετικών κόμβων σαν ομάδα, τότε ο αλγόριθμος συνεχίζει με τον κόμβο που απέχει FSON+1 hops από τον προαιρετικό που δε βρέθηκε, όπου FSON η τιμή του σχετικού πεδίου του συγκεκριμένου προαιρετικού κόμβου.

Επίσης, αν ταιριάξει το πρότυπο, δηλαδή βρεθεί ένα στιγμιότυπο επιθυμητής πληροφορίας, τότε γίνεται η εξαγωγή των δεδομένων που έχει ορίσει ο χρήστης και μετά το πρότυπο «αδειάζει» ώστε να γίνει αναζήτηση νέου ταιριάσματος μέσα στην αναπαράσταση myDOM.

Τα παραπάνω θα γίνουν καλύτερα κατανοητά με το αντιπροσωπευτικό παράδειγμα που ακολουθεί. Στην Εικόνα 28 απεικονίζεται αριστερά το πρότυπο Pattern και δεξιά το δέντρο στόχος Tree. Έστω ότι οι κόμβοι D, E, I, X, T, N, F και K του Pattern είναι υποχρεωτικοί και ο κόμβος O είναι προαιρετικός, ενώ οι F, N και K είναι αυτοί που το περιεχόμενο τους ενδιαφέρει το χρήστη. Εικονική ρίζα του

κανόνα είναι το X και πραγματική ρίζα το D. Είναι εμφανές στην εικόνα ποιος κόμβος ταιριάζει με ποιον κατά την εκτέλεση του αλγορίθμου συμφωνίας προτύπου.

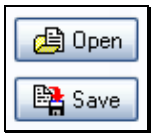


Εικόνα 28: Πρότυπο και δέντρο στόχος παραδείγματος

Αυτοματοποιημένος Τρόπος Λειτουργίας

Εφόσον ο χρήστης έχει κατασκευάσει ένα κανόνα εξαγωγής που έχει ικανοποιητική απόδοση και εξάγει τα επιθυμητά δεδομένα, έχει εξαιρετική σημασία γι' αυτόν η δυνατότητα να μπορεί να τον αποθηκεύει ώστε να τον χρησιμοποιεί κάθε φορά που τον χρειάζεται. Έτσι, δεν θα είναι αναγκασμένος να φτιάχνει τους ίδιους κανόνες ξανά και ξανά για τις ίδιες σελίδες. Για το σκοπό αυτό υλοποιήθηκε λειτουργία αποθήκευσης ενός wrapper για μελλοντική χρήση και παρέχεται η δυνατότητα εκτέλεσης του ανά πάσα στιγμή έπειτα από εντολή του χρήστη.

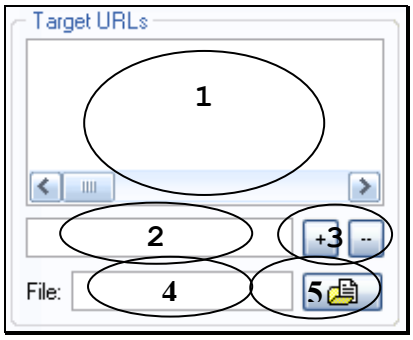
Όλες οι απαραίτητες για τον wrapper πληροφορίες αποθηκεύονται σε ένα XML αρχείο, ώστε να μπορεί ο χρήστης να «φορτώσει» και να χρησιμοποιήσει τον wrapper με τις ρυθμίσεις που είχε κάνει κατά την κατασκευή του. Τα XML αρχεία αυτά ονομάζονται *wrapper project files* και έχουν *wpf* κατάληξη, ενώ ακολουθούν τους συντακτικούς κανόνες που θέτει το DTD (wpf.dtd) που κατασκευάστηκε για την επαλήθευση της εγκυρότητας τους. Για αποθήκευση και άνοιγμα *wpf* αρχείων υπάρχουν σχετικά κουμπιά (Εικόνα 29) στην καρτέλα *Project Info*. Να σημειωθεί πως για το άνοιγμα ενός wpf αρχείου πρέπει το wpf.dtd να βρίσκεται στον ίδιο φάκελο με το wpf.



Εικόνα 29: Κουμπιά για άνοιγμα και αποθήκευση wrapper

Ένας wrapper μπορεί να εκτελεστεί για πολλά URLs. Αυτό έχει νόημα για σελίδες της ίδιας μορφής, για παράδειγμα σελίδες του ίδιου δικτυακού τόπου που έχουν διαφορετικό μεν περιεχόμενο αλλά είναι του ίδιου τύπου. Ο wrapper επισκέπτεται κάθε μία ξεχωριστά, εφαρμόζει το πρότυπο για να εντοπίσει στιγμιότυπα επιθυμητής πληροφορίας και εμφανίζει τα αποτελέσματα με ενιαίο τρόπο. Για το σκοπό αυτό, ο χρήστης μπορεί να ορίσει ως είσοδο του wrapper είτε μία λίστα από URLs είτε ένα αρχείο κειμένου που περιέχει τα URLs των πηγών στόχων.

Ο καθορισμός των πηγών στόχων επιτυγχάνεται μέσω σχετικών στοιχείων (Εικόνα 30) στην καρτέλα *Project Info*. Να σημειωθεί πως όταν ο χρήστης επισκέπτεται μία σελίδα, τότε η διεύθυνση της εισάγεται αυτόματα στη λίστα της περιοχής 1 στην Εικόνα 30, ενώ προηγουμένως έχουν απορριφθεί τυχόν προϋπάρχοντα περιεχόμενα της.



Εικόνα 30: Ορισμός πηγών στόχων ενός wrapper

Ο χρήστης με τα κουμπί '+' και τη βοήθεια του σχετικού πεδίου κειμένου, που φαίνονται στις περιοχές 3 και 2 αντίστοιχα στην Εικόνα 30, μπορεί να προσθέτει URLs. Επίσης, με επιλογή URL(s) της λίστας και το κουμπί '–' μπορεί να απομακρύνει URLs. Με το κουμπί στην περιοχή 5 της ίδιας εικόνας, ανοίγει παράθυρο διαλόγου που επιτρέπει το άνοιγμα αρχείου κειμένου με URLs και εισαγωγή των URLs που αυτό περιέχει στη λίστα. Στο πεδίο κειμένου στην περιοχή 4 απεικονίζεται το μονοπάτι του καθορισμένου αρχείου. Μάλιστα, ο χρήστης έχει τη δυνατότητα να πληκτρολογήσει απευθείας στο πεδίο αυτό το μονοπάτι (σχετικό ή

απόλυτο) του αρχείου, καθορίζοντας έτσι ως πηγές στόχους τα URLs του αρχείου. Πρέπει να σημειωθεί πως στην περίπτωση αυτή, δεν γίνεται εισαγωγή των URLs στη λίστα. Επιπλέον, να παρατηρηθεί πως για την αποθήκευση ενός wrapper, ο χρήστης πρέπει να επιλέξει έναν εκ των δύο τρόπων ορισμού των πηγών στόχων (λίστα ή αρχείο).

Για την εκτέλεση ενός αποθηκευμένου wrapper, αρκεί ο χρήστης να φορτώσει το *project* αρχείο μέσω του σχετικού κουμπιού (`Open`) και να πατήσει το κουμπί αυτόματης εκτέλεσης (Εικόνα 31). Σε περίπτωση που ο χρήστης επιθυμεί για κάποιο λόγο τη διακοπή της εκτέλεσης του κανόνα, μπορεί να πατήσει το σχετικό κουμπί (Εικόνα 15) στην περιοχή 2 της εφαρμογής.

Εικόνα 31: Κουμπί αυτόματης εκτέλεσης wrapper

Όταν φορτώνεται ένα project αρχείο, τότε τα σχετικά συστατικά του εργαλείου παίρνουν τις τιμές που καθορίζονται στα αντίστοιχα XML στοιχεία του αρχείου. Ο κανόνας εξαγωγής φορτώνεται στη δενδρική δομή προτύπου και οι διευθύνσεις των πηγών στόχων εισάγονται στην αντίστοιχη λίστα. Μάλιστα με διπλό κλικ σε οποιαδήποτε διεύθυνση της λίστας, ανοίγει η συγκεκριμένη σελίδα σε νέο παράθυρο του Internet Explorer και έτσι είναι δυνατή η επισκόπηση μίας σελίδας-στόχου. Σε περίπτωση που πρέπει να αγνοηθούν κάποιοι τύποι HTML στοιχείων, τότε ενεργοποιούνται τα σχετικά κουμπιά ελέγχου στη λίστα για τη λειτουργία απλοποίησης της myDOM αναπαράστασης. Επίσης, γίνονται οι κατάλληλες αναθέσεις τιμών στα στοιχεία για ενδεχόμενη έξοδο σε αρχείο, όπως και σε αυτά που αφορούν στην πλοήγηση σε επόμενες σελίδες ακολουθώντας '`Next`' συνδέσμους.

Ιδιαίτερο ενδιαφέρον παρουσιάζει η δυνατότητα ορισμού πολλαπλών πηγών στόχων μέσω αρχείου, καθώς επιτρέπει τη γραμμική συσχέτιση wrappers όπου τα αποτελέσματα εξόδου ενός wrapper χρησιμοποιούνται ως είσοδος από έναν άλλο. Για παράδειγμα, ο πρώτος wrapper (`w1`) εξάγει από μία ή περισσότερες σελίδες ενός δικτυακού τόπου τα `URLs` στα οποία βρίσκεται η πραγματικά επιθυμητή πληροφορία και τα αποθηκεύει σε ένα αρχείο κειμένου. Ο δεύτερος wrapper (`w2`) μπορεί να χρησιμοποιήσει ως πηγές στόχους τα `URLs` που εξήγαγε ο `w1` σε αρχείο. Έτσι, ο `w2` επισκέπτεται τις σελίδες αυτές και τελικά εξάγει τα ζητούμενα δεδομένα. Με αυτόν

τον τρόπο, υποστηρίζεται έμμεση συνεργασία wrappers, η οποία αποτελεί σημαντικό πλεονέκτημα του συστήματος.

Τέλος, αξίζει να σημειωθεί ότι το ΔEiXTo μπορεί να εκτελεστεί και από γραμμή εντολών (command line) με παράμετρο το wpf αρχείο που περιέχει όλες τις απαραίτητες πληροφορίες για τη συγκεκριμένη εκτέλεση. Έτσι καθίσταται δυνατή η συνεργασία με κάποια εφαρμογή χρονοπρογραμματισμού εργασιών (scheduler), όπως το *MS Scheduled Tasks*, και η περιοδική εκτέλεση wrappers.

Αυτόματη Υποβολή Φόρμας

Το ΔEiXTo έχει τη δυνατότητα κατά την αυτόματη εκτέλεση ενός wrapper να υποβάλλει σε ένα δικτυακό τόπο ένα ερώτημα χρήστη και έπειτα να εφαρμόζει στις σελίδες των αποτελεσμάτων τον κανόνα εξαγωγής που παρέχεται στο *project* αρχείο. Συγκεκριμένα, υποβάλλεται η φόρμα με το ερώτημα χρήστη, μεταφορτώνεται η πρώτη σελίδα αποτελεσμάτων, εξάγονται τα στιγμιότυπα επιθυμητής πληροφορίας και η διαδικασία συλλογής δεδομένων συνεχίζεται με τις επόμενες σελίδες αφού ο wrapper έχει τη δυνατότητα να ακολουθεί τους '`Next`' υπερσυνδέσμους.

Η λειτουργία αυτή είναι ιδιαίτερα χρήσιμη για εξαγωγή πληροφορίας από μηχανές αναζήτησης, ιστοχώρους ηλεκτρονικών καταστημάτων και υπηρεσίες σύγκρισης τιμών. Αρκεί ο χρήστης να προσδιορίσει το όνομα της φόρμας, το όνομα του πεδίου αναζήτησης και το ερώτημα του. Μάλιστα, τα δύο πρώτα είναι προαιρετικά. Αν ο χρήστης αφήσει για παράδειγμα το όνομα της φόρμας κενό, τότε επιλέγεται το πρώτο στοιχείο φόρμας που θα βρεθεί στη σελίδα. Τις πληροφορίες αυτές τις εισάγει ο χρήστης στα σχετικά πεδία που βρίσκονται στην καρτέλα *Project Info*, τα οποία απεικονίζονται στην Εικόνα 32.

Εικόνα 32: Στοιχεία για αυτόματη υποβολή φόρμας

Έτσι, ο χρήστης δεν είναι αναγκασμένος να δίνει ως είσοδο στο wrapper μία συγκεκριμένη διεύθυνση στόχο, στην προκειμένη περίπτωση αυτή της πρώτης σελίδας αποτελεσμάτων για το συγκεκριμένο ερώτημα. Αυτό που πρέπει να κάνει

κατά την κατασκευή ενός wrapper είναι να εισάγει, ως URL στόχο, τη διεύθυνση της αρχικής σελίδας του δικτυακού τόπου και να συμπληρώσει τα πεδία που αναφέρθηκαν για υποβολή φόρμας. Κάθε φορά λοιπόν που θέλει να εξάγει δεδομένα από αυτόν τον δικτυακό τόπο, ανεξάρτητα από το ερώτημα που θέλει να πραγματοποιήσει, εκτελεί τον ίδιο wrapper και το μόνο που χρειάζεται να αλλάξει είναι η λέξη κλειδί ή η φράση που χρησιμοποιείται για την αναζήτηση.

Διόρθωση Κανόνα Εξαγωγής

Εφόσον αποθηκευτεί ένας κανόνας, μπορεί ο χρήστης να τον χρησιμοποιήσει. Κατά βούληση. Μπορεί όμως αργότερα κάποια στιγμή, λόγω αλλαγών στη δομή της σελίδας στόχου, να πάψει ο κανόνας να είναι το ίδιο αποτελεσματικός. Πιθανό είναι επίσης και το ενδεχόμενο ο χρήστης για δικούς του λόγους να επιθυμεί μικρές αλλαγές στον κανόνα. Τίθεται λοιπόν ένα σημαντικό ζήτημα, η δυνατότητα συντήρησης (maintenance) των κανόνων που έχουν κατασκευαστεί. Επιθυμητό, φυσικά, χαρακτηριστικό είναι η δυνατότητα εύκολης τροποποίησης ενός κανόνα, ώστε να μην απαιτείται κατασκευή νέου κανόνα από το μηδέν (from scratch).

Το πρότυπο όμως δεν έχει δεδομένα και η αντίστοιχη δενδρική δομή της καρτέλας *Project Info* είναι μόνο για ανάγνωση (read-only). Ζητούμενο είναι να βρεθεί ένα στιγμιότυπο της σελίδας στόχου σύμφωνο με το πρότυπο και να τοποθετηθεί στη δενδρική δομή του κανόνα-στιγμιότυπου, ώστε να μπορεί ο χρήστης να το τροποποιήσει κατάλληλα και έτσι να βελτιώσει τον κανόνα του. Η τροποποίηση του κανόνα επιτυγχάνεται με τη λειτουργία διόρθωσης κανόνα.

Όταν ο χρήστης πατήσει το κουμπί διόρθωσης `Tune` (Εικόνα 33) στην καρτέλα *Project Info*, τότε το πρόγραμμα πλοήγησης επισκέπτεται τη σελίδα στόχο και ψάχνει στη myDOM αναπαράστασή της για πλήρες ταίριασμα (full match) με το πρότυπο, ώστε να δημιουργηθεί ένα στιγμιότυπο επιθυμητής πληροφορίας, το οποίο θα έχει δεδομένα σε όλους τους κόμβους.

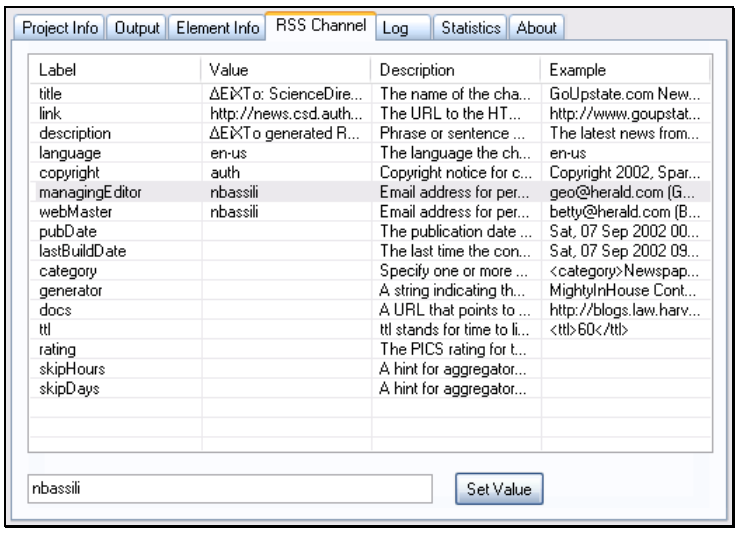
Εικόνα 33: Κουμπί Tune

Η διαδικασία αναζήτησης πλήρους ταιριάσματος σταματά είτε όταν αυτό βρεθεί είτε όταν εξαντληθούν ανεπιτυχώς όλες οι σελίδες-στόχοι. Στο στιγμιότυπο που θα εντοπιστεί, ο χρήστης μπορεί να κάνει τις αλλαγές και τις ρυθμίσεις που

επιθυμεί, ώστε να βελτιώσει την απόδοση του κανόνα. Σε περίπτωση βέβαια που υπάρχουν σοβαρές αλλαγές σε μια σελίδα, υπάρχει το ενδεχόμενο να μη βρεθεί κάποιο στιγμιότυπο που να ταιριάζει πλήρως με το πρότυπο. Αν δε βρεθεί πλήρες ταίριασμα με το πρότυπο, τότε απλά δεν παράγεται καθόλου κανόνας-στιγμιότυπο και εμφανίζεται κατάλληλο μήνυμα.

Έξοδος σε RSS αρχείο

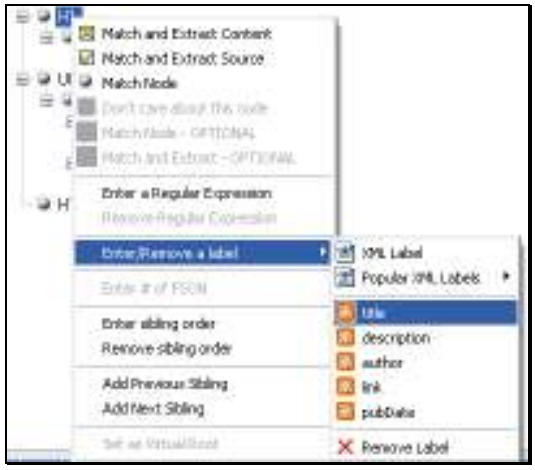
Το ΔEiXTo έχει τη δυνατότητα να παράγει έξοδο σε *RSS* αρχείο. Τα *item* στοιχεία του *channel* δημιουργούνται από τα δεδομένα που εξάγονται από τα στιγμιότυπα επιθυμητής πληροφορίας που εντοπίστηκαν. Στην καρτέλα *RSS Channel* στην περιοχή 5 της εφαρμογής ο χρήστης μπορεί να καθορίσει τις τιμές των υπό-στοιχείων (sub-elements) του *channel* στοιχείου του RSS αρχείου εξόδου (Εικόνα 34). Να σημειωθεί ότι στο υπο-στοιχείο *title* ανατίθεται αυτόματα η τιμή "ΔEiXTo: τίτλος της προβαλλόμενης ιστοσελίδας".



Εικόνα 34: Υπό-στοιχεία του channel στοιχείου του RSS αρχείου εξόδου

Ο χρήστης μπορεί να αναθέσει RSS ετικέτα σε κάθε κόμβο του κανόνα εξαγωγής, όπως φαίνεται στην Εικόνα 35. Μάλιστα, έχει την ευχέρεια να επιλέξει μεταξύ των RSS στοιχείων: *title*, *author*, *description*, *link* και *pubDate*. Έτσι, αν ο χρήστης θέσει ως τύπο αρχείου εξόδου το RSS και εκτελέσει ένα κανόνα, τότε σε αυτό το αρχείο εξάγονται τα δεδομένα των μεταβλητών εξόδου που έχουν RSS ετικέτες. Σε περίπτωση που ο χρήστης δεν έχει αναθέσει ετικέτα link σε κάποιον από τους κόμβους που εξάγονται, τότε προστίθεται αυτόματα στοιχείο link σε κάθε στοιχείο *item* του αρχείου, το οποίο έχει ως τιμή τη διεύθυνση της ιστοσελίδας από

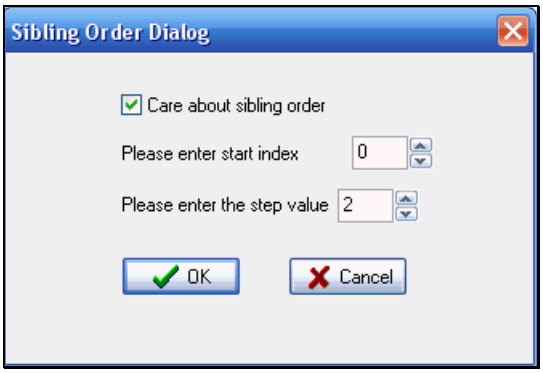
την οποία εξάγεται το στιγμιότυπο επιθυμητής πληροφορίας. Τέλος, για την εκτέλεση ενός κανόνα εξαγωγής που παράγει RSS έξοδο, απαιτείται ο κανόνας να περιλαμβάνει κόμβο που βρίσκεται σε κατάσταση checked ή checked_implied και έχει ετικέτα *RssTitle* ή *RssDescription*.



Εικόνα 35: Ανάθεση RSS ετικέτας σε κόμβο κανόνα

Τάξη Κόμβου

Σύμφωνα με την έως τώρα περιγραφή, ο αλγόριθμος συμφωνίας προτύπου στηρίζεται στη σειρά εμφάνισης των κόμβων αλλά όχι στην τάξη (sibling order) τους. Ωστόσο, σε κάποιες περιπτώσεις είναι χρήσιμο να μπορεί ο χρήστης να ορίσει την τάξη ενός κόμβου του κανόνα. Αυτό μπορεί να γίνει με σχετική επιλογή στο τοπικό μενού του κόμβου. Στο παράθυρο που ανοίγει (Εικόνα 36), ο χρήστης έχει τη δυνατότητα να διατυπώσει μαθηματικές εκφράσεις του τύπου `K*i+C`, όπου το `C` καθορίζει την αρχή (start index), `K` είναι το βήμα (step) και `i` αριθμός μεγαλύτερος ή ίσος (≥) του 0.



Εικόνα 36: Παράθυρο διαλόγου για ορισμό τάξης κόμβου από το χρήστη

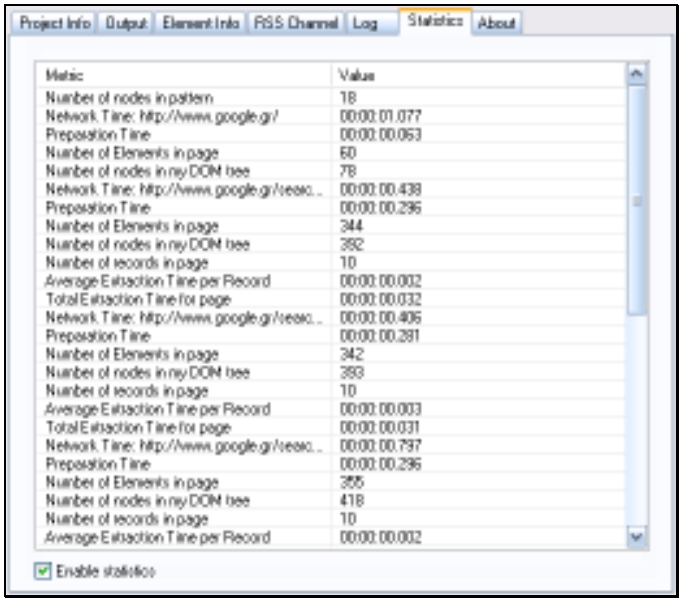
Να σημειωθεί πως η τάξη του πρώτου θυγατρικού κόμβου είναι 0. Αν κάποιος επιθυμεί σταθερή τιμή `N` για την τάξη, τότε απλά θα πρέπει να δώσει τιμή 0 στο βήμα

και N στην αρχή. Έτσι, αν για παράδειγμα ο χρήστης θέλει ένας κόμβος του προτύπου να ταιριάξει με myDOM κόμβους τάξης 0,2,4,6,.., πρέπει να ορίσει ως αρχή το 0 και βήμα το 2. Μία ενδεικτική εφαρμογή της λειτουργίας αυτής θα ήταν για παράδειγμα η εξαγωγή των άρτιων ή περιττών αποτελεσμάτων από μία αναζήτηση χρήστη.

Στατιστικά

Στην καρτέλα *Statistics* δίνονται κάποιες τιμές και μετρήσεις σχετικές με την εκτέλεση ενός wrapper και την απόδοση του συστήματος (Εικόνα 37). Συγκεκριμένα, αν το κουμπί ελέγχου της λειτουργίας καταγραφής στατιστικών, που βρίσκεται στην ίδια καρτέλα, είναι ενεργοποιημένο, τότε καταγράφονται οι εξής μετρικές:

- χρόνος δικτύου (network time): χρόνος που χρειάζεται για τη μεταφόρτωση μιας σελίδας και την πλήρη προβολή της στο πρόγραμμα πλοήγησης.
- χρόνος προετοιμασίας (preparation time): χρόνος που απαιτείται για κατασκευή της δενδροειδούς myDOM αναπαράστασης μιας σελίδας και των απαραίτητων δομών δεδομένων.
- πλήθος HTML στοιχείων σε μια σελίδα.
- αριθμός κόμβων στη myDOM αναπαράσταση μιας σελίδας.
- αριθμός κόμβων του κανόνα εξαγωγής.
- πλήθος στιγμιότυπων που εντοπίστηκαν σε μια σελίδα.
- συνολικός χρόνος εξαγωγής των επιθυμητών δεδομένων μιας σελίδας.
- μέσος χρόνος εξαγωγής για κάθε στιγμιότυπο επιθυμητής πληροφορίας.



Εικόνα 37: Στατιστικά για εκτέλεση ενδεικτικού wrapper

Παράρτημα

Κανονικές Εκφράσεις

Για την ανάπτυξη του ΔΕiΧΤο, έγινε χρήση μίας εύχρηστης αλλά και ισχυρής βιβλιοθήκης κανονικών εκφράσεων (regular expressions) με όνομα *TRegExpr*, δημιουργός της οποίας είναι ο *Andrey V. Sorokin*.

Οι *κανονικές εκφράσεις* είναι σχεδόν μια γλώσσα από μόνες τους. Είναι μία τυπική μέθοδος περιγραφής προτύπων αλφαριθμητικών (strings). Αν το πρότυπο που έχει καθοριστεί βρεθεί οπουδήποτε μέσα σε μία συμβολοσειρά στόχο, τότε υπάρχει συμφωνία (match). Άρα, οι κανονικές εκφράσεις περιγράφουν με ακρίβεια ένα σύνολο από αλφαριθμητικά, σύμφωνα με συγκεκριμένους συντακτικούς κανόνες. Χρησιμοποιούνται από πολλά προγράμματα επεξεργασίας κειμένου και άλλα βοηθητικά προγράμματα για αναζήτηση και χειρισμό κειμένου. Για τις κανονικές εκφράσεις έχουν γραφτεί ολόκληρα βιβλία, ωστόσο παρακάτω θα γίνει μια σύντομη αλλά περιεκτική περιγραφή των βασικότερων χαρακτηριστικών τους.

Οι περισσότεροι χαρακτήρες σε μία κανονική έκφραση αναπαριστούν τον εαυτό τους (literals). Οι χαρακτήρες που αποτελούν εξαίρεση ονομάζονται μεταχαρακτήρες και αλλάζουν τη συμπεριφορά της συμφωνίας προτύπου. Οι μεταχαρακτήρες είναι οι εξής:

```
^ $ ( ) \ | @ [ { ? . + *
```

Αν ένας μεταχαρακτήρας πρέπει να αντιμετωπιστεί με την κυριολεκτική του τιμή, τότε πρέπει να μπει πριν από αυτόν μία ανάστροφη κάθετος (backslash) ώστε να χάσει την ειδική σημασία του. Επίσης, υπάρχουν ορισμένοι ειδικοί χαρακτήρες οι οποίοι προσδιορίζονται με τη βοήθεια της ανάστροφης καθέτου, όπως για παράδειγμα ο `'\n'` που ορίζει μία νέα γραμμή (newline) και ο `'\t'` που ορίζει ένα στηλοθέτη (tab).

Οι κανονικές εκφράσεις συμμορφώνονται σε ορισμένους κανόνες. Αυτοί είναι οι παρακάτω:

- Το ταίριασμα προτύπου γίνεται από τα αριστερά της συμβολοσειράς στόχου προς τα δεξιά.
- Η διαδικασία ταιριάσματος προτύπου επιστρέφει αληθές (true) αν και μόνο αν ταιριάξει όλο το πρότυπο με τη συμβολοσειρά στόχου.
- Το πρώτο ταίριασμα που γίνεται είναι το αριστερότερο μέσα στη συμβολοσειρά στόχου. Οι κανονικές εκφράσεις δεν εγκαταλείπουν ένα καλό ταίριασμα ψάχνοντας για ένα άλλο, που πιθανώς να βρίσκεται παρακάτω. Ωστόσο, λαμβάνεται το μεγαλύτερο δυνατό πρώτο ταίριασμα, καθώς οι κανονικές εκφράσεις είναι «άπληστες» (greedy), με την έννοια ότι προσπαθούν να συμφωνήσουν όσο το δυνατό περισσότερα πράγματα.

Μία συνηθισμένη πρακτική στις κανονικές εκφράσεις είναι να ζητάει κάποιος ταίριασμα «οποιουδήποτε από δεδομένους χαρακτήρες». Αυτό επιτυγχάνεται μέσω της κλάσης χαρακτήρων. Για τη δημιουργία μίας κλάσης αρκεί η τοποθέτηση μίας λίστας από χαρακτήρες μέσα σε αγκύλες []. Οι χαρακτήρες σε μία κλάση χαρακτήρων θεωρούνται ως ένας μόνο χαρακτήρας, ο οποίος ταιριάζει με οποιονδήποτε χαρακτήρα που περιλαμβάνεται στη λίστα. Για παράδειγμα, η έκφραση `foob[aeiou]r` ταιριάζει με τα `'foobar'`, `'foober'` αλλά όχι με τα `'foobbr'`, `'foobcr'` κ.τ.λ. Αν ο πρώτος χαρακτήρας της κλάσης είναι `"^"`, τότε η κλάση συμφωνεί με κάθε χαρακτήρα που δεν βρίσκεται μέσα στην κλάση. Σε μία λίστα ο χαρακτήρας `'-'` χρησιμοποιείται για να ορίσει μία περιοχή (range) χαρακτήρων, π.χ. η κλάση `[a-e]` συμφωνεί οποιοδήποτε από τους `a`, `b`, `c`, `d` ή `e`.

Μέχρι τώρα, σε όλα τα παραδείγματα που δόθηκαν, όλοι οι χαρακτήρες μέσα σε πρότυπα, είτε πρόκειται για χαρακτήρες κειμένου, είτε πρόκειται για μεταχαρακτήρες, είχαν μια σχέση ένα προς ένα με χαρακτήρες στη συμβολοσειρά στόχου με την οποία προσπαθούσαν να συμφωνήσουν. Ένας ποσοτικοποιητής (quantifier) είναι ένα είδος τελεστή που «λέει» στην κανονική έκφραση πόσες συνεχόμενες εμφανίσεις ενός πράγματος να ταιριάξει και μπορεί να τοποθετηθεί μετά από ένα χαρακτήρα. Ο απλούστερος τέτοιος τελεστής είναι ο μεταχαρακτήρας `+`. Το `+` κάνει τον προηγούμενο χαρακτήρα να συμφωνήσει τουλάχιστον μια φορά ή όσες φορές μπορεί και να συνεχίσει αν υπάρχει μια έκφραση που συμφωνεί. Έτσι, το `ab+c` θα συμφωνήσει με το `deabcie` και το `wabbbcr` αλλά όχι με το `fractal` (λείπει το `b`) και το `doabbf` (λείπει το `c`). Ο μεταχαρακτήρας `*` είναι παρόμοιος με τον `+`, αλλά κάνει τον μεταχαρακτήρα του οποίου έπεται να συμφωνεί μηδέν ή περισσότερες

φορές. Έτσι, το `car*t` θα συμφωνήσει με το `carted` και το `cat` αλλά όχι με το `carrot` (το `o` διακόπτει το πρότυπο) και το `carl` (το `t` μέσα στο πρότυπο είναι υποχρεωτικό). Ένας άλλος μεταχαρακτήρας είναι ο `?` ο οποίος κάνει τον προηγούμενο χαρακτήρα να συμφωνήσει μηδέν ή μία φορά (όχι περισσότερες). Έτσι, το πρότυπο `b?all` προκαλεί τη συμφωνία ενός c αν υπάρχει. Αλλιώς, δεν υπάρχει πρόβλημα. Κατόπιν ακολουθείται από ένα `a`, `l` και `l`. Στην ουσία, αυτό το πρότυπο συμφωνεί με κάθε συμβολοσειρά που περιέχει `all`, αρκεί αμέσως πριν από αυτό να προηγείται το πολύ ένα `b`. Δηλαδή, το πρότυπο δε συμφωνεί με το `bball`.

Οι κανονικές εκφράσεις παρέχουν επίσης τη δυνατότητα στο χρήστη να συμφωνήσει ακριβώς όσες εμφανίσεις θέλει χρησιμοποιώντας τα άγκιστρα {}. Ο ποσοτικοποιητής με άγκιστρα έχει τη μορφή `string{n,m}`, όπου `n` ο ελάχιστος αριθμός συμφωνιών, `m` ο μέγιστος αριθμός συμφωνιών και `string` ο χαρακτήρας ή η ομάδα χαρακτήρων που προσπαθεί να ποσοτικοποιήσει. Μπορεί να παραλειφθεί το `n` ή το `m`, αλλά όχι και τα δύο. Έτσι το `x{5,10}` συμφωνεί όταν το `x` εμφανίζεται τουλάχιστον 5 φορές αλλά όχι περισσότερες από 10, το `x{7,}` συμφωνεί όταν το `x` εμφανίζεται τουλάχιστον 7 φορές και το `x{4}` συμφωνεί όταν το `x` εμφανίζεται ακριβώς 4 φορές. Επίσης, ένας συνηθισμένος ιδιωματισμός στις κανονικές εκφράσεις είναι το `.*` που χρησιμοποιείται για να ταιριάξει οτιδήποτε (συνήθως, οτιδήποτε ανάμεσα σε δύο άλλα πράγματα που ενδιαφέρουν το χρήστη).

Επιπλέον, υπάρχουν κάποιες συντομεύσεις για ορισμένες κλάσεις χαρακτήρων που χρησιμοποιούνται συχνά. Αυτές συμβολίζονται με μία ανάστροφη κάθετο και ένα μη μεταχαρακτήρα. Ακολουθούν ειδικές κλάσεις χαρακτήρων:

`\w`	ένας χαρακτήρας λέξης. Ίδιο με `[a-zA-Z0-9_]`
`\W`	η άρνηση του `\w`
`\d`	ένα ψηφίο του δεκαδικού συστήματος. Ίδιο με `[0-9]`
`\D`	η άρνηση του `\d`
`\s`	ένας λευκός χαρακτήρας, ίδιο με `[ \t\n\r\f]`
`\S`	η άρνηση του `\s`

Έτσι, το `\d+` θα συμφωνήσει με το `134` αλλά όχι με το `abc`. Επίσης, το `\w+\s*\D` θα συμφωνήσει με το "`A_5 b`" αλλά όχι με το "`a 3`" (το 3 είναι δεκαδικό ψηφίο και δε συμφωνεί με το `\D`).

Οι κανονικές εκφράσεις δίνουν ακόμα τη δυνατότητα εύρεσης συνόλου προτύπων. Ο μηχανισμός αυτός ονομάζεται εναλλακτική συμφωνία (alternation) και

εμφανίζεται σε μια κανονική έκφραση όταν πιθανές συμφωνίες χωρίζονται με ένα χαρακτήρα |. Ακόμα παρέχεται η δυνατότητα ομαδοποίησης τμημάτων του προτύπου με παρενθέσεις, (). Για παράδειγμα η έκφραση `(c|p|s)at` συμφωνεί τις λέξεις `cat`, `pat` και `sat`. Αξίζει να σημειωθεί επίσης, πως οι παρενθέσεις σε κανονικές εκφράσεις επιτρέπουν, εφόσον επιτυγχάνεται ταίριασμα, την απομόνωση και αποθήκευση σε ειδικές μεταβλητές, του τμήματος της συμβολοσειράς στόχου που συμφώνησε με την κάθε έκφραση σε παρενθέσεις.

Οι δύο τελευταίοι μεταχαρακτήρες είναι οι άγκυρες (anchors), οι οποίες καθορίζουν το ακριβές σημείο στο οποίο θα γίνει η αναζήτηση για το πρότυπο- στην αρχή ή στο τέλος μιας συμβολοσειράς. Η πρώτη άγκυρα είναι το `"^"`. Ο χαρακτήρας αυτός στην αρχή μιας κανονικής έκφρασης υποχρεώνει την έκφραση να ψάξει για συμφωνία μόνο στην αρχή μίας συμβολοσειράς. Για παράδειγμα το `^dog` συμφωνεί τη λέξη `dog` μόνο αν αυτή εμφανίζεται στην αρχή μιας συμβολοσειράς. Η δεύτερη άγκυρα είναι το `$`, που στο τέλος μίας κανονικής έκφρασης υποχρεώνει να γίνει το ταίριασμα μόνο στο τέλος μίας συμβολοσειράς.

Τέλος, εκτός από την εύρεση προτύπων σε συμβολοσειρές παρέχεται και η δυνατότητα αντικατάστασης των δεδομένων που συμφωνούν με το πρότυπο. Ακόμα υπάρχουν ορισμένοι τροποποιητές (modifiers) που αλλάζουν τη συμπεριφορά των κανονικών εκφράσεων, όπως το `i` που καθιστά τη συμφωνία προτύπου μη ευαίσθητη στον τύπο των χαρακτήρων (πεζά ή κεφαλαία) και το `g` που αναγκάζει τη συμφωνία να εκτελείται επαναληπτικά μέσα σε όλη τη συμβολοσειρά, όπου το κάθε ταίριασμα γίνεται αρχίζοντας αμέσως μετά το τελευταία ταίριασμα.

Οι κανονικές εκφράσεις αποτελούν ένα πραγματικά πολύ ισχυρό εργαλείο για χειρισμό αλφαριθμητικών. Αν αναλογιστεί κανείς ότι μία ιστοσελίδα αποτελεί καταρχήν ένα αλφαριθμητικό, αντιλαμβάνεται τη σημασία αυτής της μεθόδου και της πληθώρας των εφαρμογών και των δυνατοτήτων της στον τομέα της εξαγωγής περιεχομένου από ιστοσελίδες.