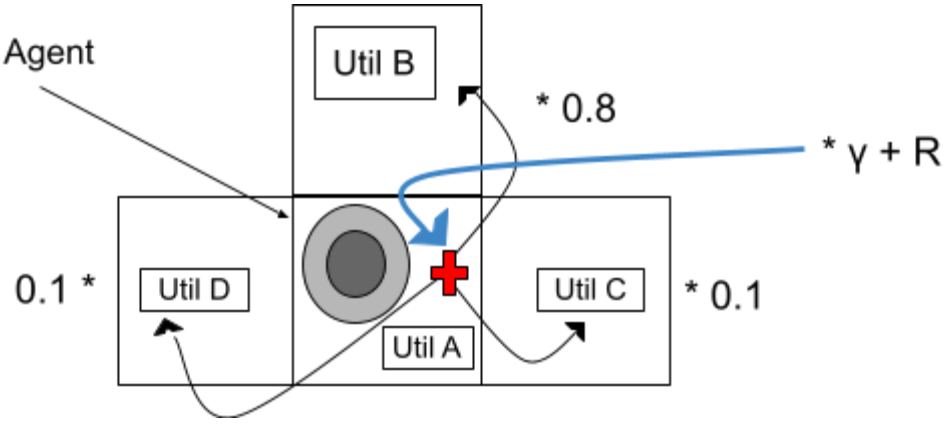
Τεχνική περιγραφή λειτουργίας πράκτορα για την εύρεση χρησιμοτήτων καταστάσεων και επίλυσης του A*

Θεμελιώδεις γνώσεις Τεχνητής Νοημοσύνης
Ματθαίος Ζηδιανάκης

Περιγραφή προσέγγισης

Το μοντέλο που χρησιμοποιήθηκε για την εύρεση των χρησιμοτήτων σε κάθε κατάσταση βασίζεται στην εξίσωση Bellman, με την οποία ο πράκτορας εξετάζει όλες τις καταστάσεις επαναληπτικά χωρίς να κινείται πραγματικά στο χώρο και βρίσκει το προεξοφλημένο άθροισμα των ανταμοιβών για κάθε κατάσταση, λαμβάνοντας υπόψη το συντελεστή προεξόφλησης *γ*. Αυτό σημαίνει ότι για κάθε εξέταση μιας κατάστασης και σε κάθε επανάληψη εξέτασης όλων των καταστάσεων, προστίθεται η ανταμοιβή που θα πάρει εκτελώντας μια κίνηση στην επόμενη πιθανή (λόγω μη αιτιοκρατικότητας) κατάσταση. Στο παρακάτω σχήμα, φαίνεται η λειτουργία αυτή για έναν κόμβο A που εξετάζεται σε μια από τις επαναλήψεις, για τον οποίο ισχύει η σχέση $UtilA = R + \gamma \cdot (0.8 \cdot UtilB + 0.1 \cdot UtilC + 0.1 \cdot UtilD)$:

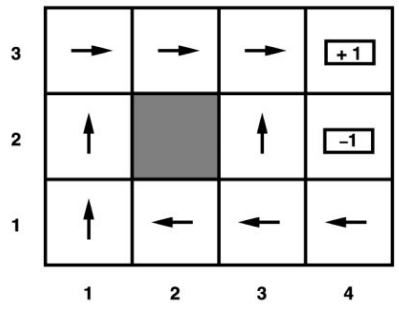


Παράδειγμα υπολογισμού χρησιμότητας για μια πιθανή κατάσταση A

Όπως στο παράδειγμα, κάθε αντίστοιχος κόμβο *Util B, Util C* και *Util D* αναφέρεται στην ανάθεση της τιμής τους από την προηγούμενη επανάληψη εξέτασης όλων των κόμβων. Έπειτα υλοποιείται ο greedy αλγόριθμος A* για την εύρεση του καλύτερου μονοπατιού μέχρι τον τελικό κόμβο +1.

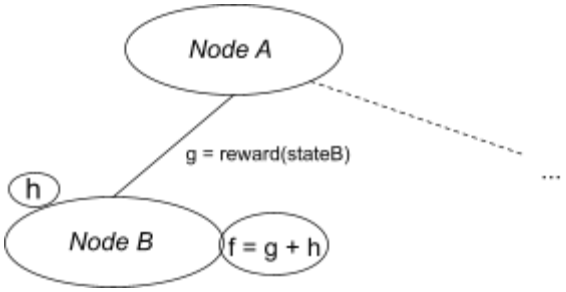
Περιγραφή αλγορίθμου

Για την ανάπτυξη του αλγορίθμου εύρεσης χρησιμοτήτων χρησιμοποιήθηκε αρχικά η δοθείσα πολιτική για την εκτέλεση των ενεργειών ανά κατάσταση:



Επίσης, ο αλγόριθμος προσφέρει την επιλογή για εύρεση από τον πράκτορα της καλύτερης πολιτικής που μπορεί να επιλέξει, βρίσκοντας τη μέγιστη εκτιμώμενη χρησιμότητα ελέγχοντας κάθε δυνατή επιλογή κίνησης στο χώρο. Ο αλγόριθμος εκτελεί την παραπάνω λειτουργία μέχρις ότου να επιτευχθεί σύγκλιση σύμφωνα με το συντελεστή $e = 0.0001$. Σε κάθε περίπτωση τρόπου επιλογής πολιτικής, για κάθε κόμβο υπολογίζεται η χρησιμότητα λαμβάνοντας υπόψη το εμπόδιο και τους "τοίχους" (όρια του χώρου). Συγκεκριμένα, όταν συναντά και "πέφτει" σε κάποιο από τα εμπόδια, μένει στη θέση που βρίσκονταν. Τελικά, εκτυπώνονται ο πίνακας με τις αναμενόμενες χρησιμότητες για κάθε κόμβο και ο πίνακας που περιέχει τις κινήσεις του πράκτορα για να επιτύχει τη βέλτιστη αναμενόμενη χρησιμότητα.

Μετά τους παραπάνω υπολογισμούς, χρησιμοποιούνται οι υπολογισμένες αναμενόμενες χρησιμότητες ως ευριστικές για τον αλγόριθμο A*, ώστε να βρει το βέλτιστο μονοπάτι μέχρι τον στόχο +1. Η συγκεκριμένη υλοποίηση του A* τροποποιήθηκε ώστε να υπολογίζεται το μονοπάτι με τη μέγιστη συνολική χρησιμότητα (αντί το ελάχιστο κόστος), δηλαδή κάθε εκτιμώμενη χρησιμότητα δείχνει πόσο απέχει από το στόχο, ώστε να μπορεί να επιλέξει την καλύτερη ακολουθία κινήσεων. Στο τέλος, εκτυπώνεται το μονοπάτι που ακολούθησε για το στόχο +1, καθώς και η θέση του στόχου. Για το κόστος *g* το οποίο θα προστεθεί στην τιμή της ευριστικής *h* για την εύρεση του συνολικού κόστους μονοπατιού *f*, χρησιμοποιείται η ανταμοιβή που θα πάρει εκτελώντας μια ενέργεια, όπως φαίνεται παρακάτω:



Παράδειγμα υπολογισμού του συνολικού κόστους από τον κόμβο A στον B

Αποτίμηση αποτελεσμάτων

Χρησιμοποιώντας την παραπάνω πολιτική του σχήματος, ο αλγόριθμος για $\gamma = 0.9$, $\gamma = 0.6$ και $\gamma = 0.2$ παράγει τα παρακάτω αποτελέσματα:

- $\gamma = 0.9$

```
Calculating utilities for gamma = 0.9...
Finished!
Calculated expected utilities for each state
```

```
-------------------------------------------

[[ 0.50941558  0.64958636  0.79536224  1.        ]
 [ 0.39851121  0.          0.48644046 -1.        ]
 [ 0.2918707   0.20749515  0.16832337 -0.00968258]]

Policy for each state
------------------------

[['->' '->' '->' '0.']
 ['^' '8.' '^' '0.']
 ['^' '<-' '<-' '<-']]
```

- $\gamma = 0.6$

```
Calculating utilities for gamma = 0.6...

Finished!

Calculated expected utilities for each state
-------------------------------------------

[[ 0.06646175  0.21463819  0.47683669  1.        ]
 [-0.00920305  0.          0.13710809 -1.        ]
 [-0.05196544 -0.0738045  -0.07149806 -0.14290881]]

Policy for each state
------------------------

[['->' '->' '->' '0.']
 ['^' '8.' '^' '0.']
 ['^' '<-' '<-' '<-']]
```

- $\gamma = 0.2$

```
Calculating utilities for gamma = 0.2...

Finished!

Calculated expected utilities for each state
-------------------------------------------

[[-0.04534272 -0.02140672  0.12160272  1.        ]
 [-0.04926464  0.         -0.04137552 -1.        ]
 [-0.04992    -0.04992    -0.04975472 -0.0693    ]]

Policy for each state
------------------------

[['->' '->' '->' '0.']
 ['^' '8.' '^' '0.']
 ['^' '<-' '<-' '<-']]
```

Όπως βλέπουμε παραπάνω, ο πρώτος πίνακας αντιπροσωπεύει το περιβάλλον που βρίσκεται ο πράκτορας και κάθε αριθμός που είναι ακέραιο μηδέν, 1 ή -1 αντιπροσωπεύει το εμπόδιο στο κέντρο και τους στόχους αντίστοιχα, αφού δεν υπολογίζονται εκτιμώμενες χρησιμότητες για τους κόμβους αυτούς. Οι υπόλοιποι αριθμοί απεικονίζουν τις εκτιμώμενες χρησιμότητες του πράκτορα. Στον τελευταίο πίνακα φαίνονται οι εκτιμώμενες κινήσεις του πράκτορα στο περιβάλλον για να επιτύχουν τις παραπάνω χρησιμότητες, όπου οι τιμές 8 και 0

αντιπροσωπεύουν το εμπόδιο και τους στόχους αντίστοιχα, ενώ τα σύμβολα ^, *v*, ->, <- αντιστοιχίζονται με τις κινήσεις “πάνω”, “κάτω”, “δεξιά” και “αριστερά” αντίστοιχα.

Παρακάτω παρουσιάζονται τα αποτελέσματα χρησιμοποιώντας την επιλογή εύρεσης της καλύτερης πολιτικής κινήσεων από το πράκτορα:

- $\gamma = 0.9$

```
Calculating utilities for gamma = 0.9...

Finished!

Calculated expected utilities for each state
-------------------------------------------

[[ 0.50941508  0.64958635  0.79536224  1.        ]
 [ 0.39850958  0.          0.48644045 -1.        ]
 [ 0.29646181  0.25395726  0.34478721  0.12994003]]

Policy for each state
---------------------

[['->' '->' '->' '0.']
 ['^' '8.' '^' '0.']
 ['^' '->' '^' '<-']]
```

- $\gamma = 0.6$

```
Calculating utilities for gamma = 0.6...

Finished!

Calculated expected utilities for each state
-------------------------------------------

[[ 0.06646019  0.2146381   0.47683667  1.        ]
 [-0.00920788  0.          0.13710804 -1.        ]
 [-0.0495172  -0.0352962   0.01862407 -0.08451019]]

Policy for each state
---------------------

[['->' '->' '->' '0.']
 ['^' '8.' '^' '0.']
 ['^' '->' '^' 'v']]
```

- $\gamma = 0.2$

```
Calculating utilities for gamma = 0.2...

Finished!

Calculated expected utilities for each state
-------------------------------------------

[[-0.04533248 -0.0214016   0.12160448  1.        ]
 [-0.04926464  0.         -0.04137152 -1.        ]
 [-0.04992    -0.04975616 -0.04860416 -0.04989952]]

Policy for each state
---------------------
```

```
[['->' '->' '->' '0.']
 ['^' '8.' '^' '0.']
 ['<-' '->' '^' 'v']]
```

Παρατηρούμε ότι σε αυτήν την περίπτωση οι βέλτιστες κινήσεις στο κάτω μέρος είναι αυτές που τροποποιούνται.

Όσον αφορά τα αποτελέσματα του A*, ο αλγόριθμος θα ακολουθήσει σχεδόν σε όλες τις περιπτώσεις την αντίστοιχη πολιτική που υπολόγισε ή πήρε ως δεδομένη ως είσοδο και η αφετηρία για τον A* είναι στον κάτω αριστερά κόμβο:

```
Step through 2,0
Step through 1,0
Step through 0,0
Step through 0,1
Step through 0,2
Step through 0,3
Goal node found!

  Position: 0,3
```

Η μόνη περίπτωση που δεν ακολουθεί την πολιτική ως βέλτιστο μονοπάτι είναι η τελευταία, δηλαδή στην περίπτωση όπου $\gamma = 0.2$, γεγονός το οποίο μπορεί να διερευνηθεί περαιτέρω.

Εξετάζοντας την ευριστική του A*, μπορούμε να δούμε ότι είναι αποδεκτή καθώς δεν υπερεκτιμά σε κάθε κατάσταση το πραγματικό κόστος για να φτάσει στον επιθυμητό στόχο, αφού κάθε ανταμοιβή για μια κίνηση είναι πολλαπλασιασμένη με $\gamma < 1$, οπότε το άθροισμά τους θα είναι μικρότερο από το άθροισμα των πραγματικών ανταμοιβών που θα αποκτήσει ο πράκτορας, αν ακολουθούσε αιτιοκρατικά την πολιτική που υπολόγισε.

Επομένως, μια αιτία για την παραπάνω περίπτωση μπορεί να είναι ότι στην κάτω αριστερά κατάσταση ο πράκτορας πραγματοποιώντας την καλύτερη δυνατή ενέργεια που βρήκε, θα καταλήξει να μην κινηθεί στο χώρο, αποκτώντας μια ανταμοιβή R=-0.04, εκτός αν λόγω στοχαστικότητας αλλάξει πορεία και κινηθεί προς τα πάνω.