Отчет за работата по докторантурата на Димитър Добрев на тема „Изкуствен интелект – дефиниция, реализация и последствия“

13 януари 2023 г.

Димитър Добрев

d@dobrev.com

Институт по математика и информатика
Българска академия на науките

Език за описание на светове

Класът на изчислимите функции е един от основните обекти на изследване в областта на математическата логика. Макар изчислимите функции да са много важни за логиците, на тях винаги им е било тясно в това множество и те винаги са се опитвали да излязат от него.

Например при Тюринговите и при номерационните степени на неразрешимост се добавя оракул и по този начин се излиза от множеството на изчислимите функции. В тази дисертация също ще използваме оракули, но тези оракули ще са по-различни.

1. Оракулът „враг“. Това е агент, който играе срещу нас и винаги избира възможно най-лошия за нас ход.
2. Оракулът „случайност“. Този оракул можем да си го мислим като зар, който връща няколко възможности. Всяка от тези възможности си има вероятност.
3. Оракулът „шум“. Това е агент, който е константа, но константа с шум.

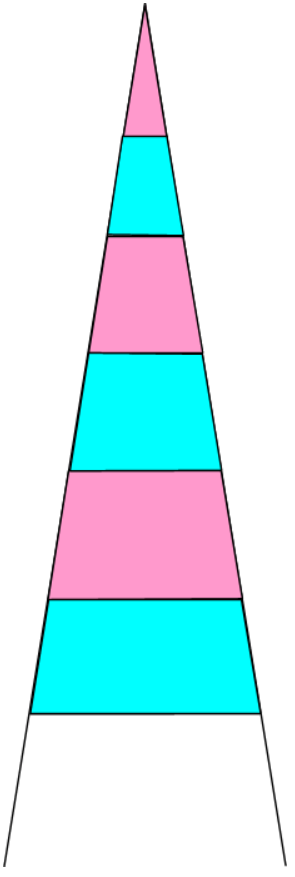
Какво наричаме „свят“?

Свят ще бъде стратегия на света в дървото на всички възможности. Светът го представяме, чрез функция:

$$f: \mathbb{N} \times \Sigma \rightarrow \Omega \times \mathbb{N}$$

Σ – множеството на действията.
Ω – множеството на наблюденията.
$\mathbb{N}$ – множеството на вътрешните състояния на света.

Дървото на всички възможности е нещо като двоично дърво с тази разлика, че разклоненията не са 2, а са $|\Sigma|$ или $|\Omega|$ съответно.



Формална дефиниция на ИИ

В първата част на дисертацията при помощта на „език за описание на светове“ дефинираме най-добрата стратегия на агента. След това даваме формална дефиниция на ИИ.

ИИ е изчислима стратегия, която е достатъчно близо до най-добрата.

В дисертацията се изказва предположението, че формалната дефиниция на ИИ не зависи от езика за описание на светове, който сме използвали. Това предположение не е доказано и вероятно въобще не може да бъде доказано.

Програма удовлетворяваща дефиницията на ИИ

На базата на избрания език за описание на светове се прави програма, която удовлетворява дефиницията на ИИ. Тази програма е толкова неефективна, че тя на практика се нуждае от безкрайно бърз компютър, макар че теоретично това е програма, завършваща след краен брой стъпки.

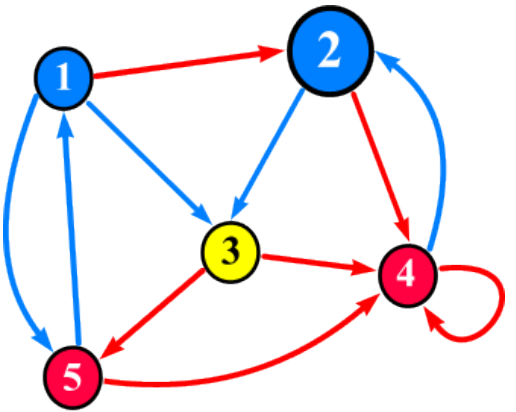
Твърдим, че тази програма е достатъчно близо до най-добрата стратегия на агента.

Понятието „достатъчно близо“ не е формално, но ние разглеждаме множество от алгоритми, които са безкрайно близо, което вече е формално. (Безкрайно близо означава, че за всяко ε съществува стойност на параметъра h такава, че при тази стойност на параметъра h съответният алгоритъм е на разстояние, по-малко от ε от най-добрата стратегия.)

Конкретен език за описание на света

Във втората част на дисертацията предлагаме един конкретен език за описание на светове, който е базиран върху събитийните модели.

Тези модели приличат на крайни автомати, но стрелките не са по букви, а по събития. Състоянията не са еднакви и затова имаме няколко вида състояния. (При крайните автомати състоянията са само два вида.)



Предимство на събитийните модели е, че те са устойчиви към модификация, докато компютърните програми не са.

Дефиниция на понятието „алгоритъм“

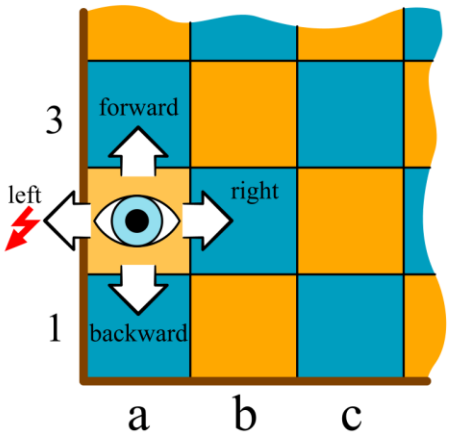
Алгоритъмът го дефинираме като последователност от действия, които се извършват в конкретен свят. Резултатът от изпълнението на алгоритъма зависи от това в кой свят го изпълняваме.

Например машината на Тюринг се състои от глава и безкрайна лента. Според нашата дефиниция алгоритъмът е само главата, а безкрайната лента е част от света, в който изпълняваме алгоритъма. Резултата от изпълнението на този алгоритъм е изчислима функция, но ако пуснем същата глава в друг свят (например върху крайна лента) ще получим друг резултат.

Алгоритъмът се описва чрез събитиен модел.

Конкретен свят

Конкретният свят е светът на играта шах, но в този свят има една интересна особеност. Интересното в случая е, че една стъпка в този свят не е местенето на една фигура. За да се премести една фигура, трябва да се направят много стъпки. Агентът „вижда“ само едно от квадратчетата на таблото. За да премести фигура, той трябва да премести погледа си върху тази фигура, да я вдигне, да премести погледа си върху новата позиция и да спусне там вдигнатата фигура. По този начин местенето на една фигура е изпълнение на алгоритъм, състоящ се от много стъпки.

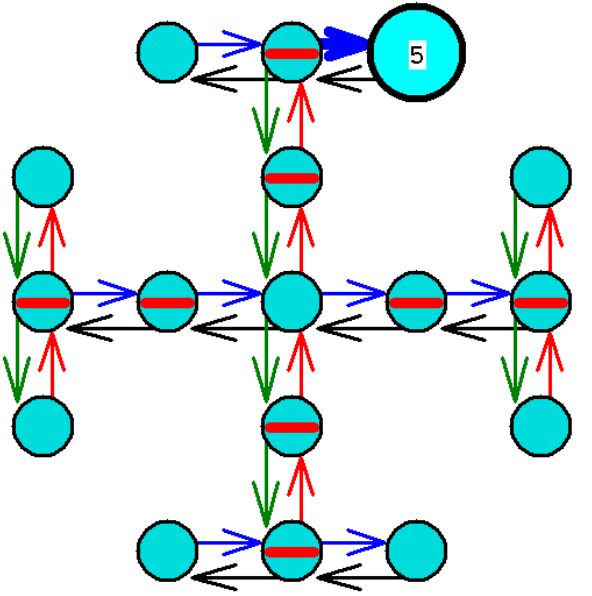


Голяма стъпка

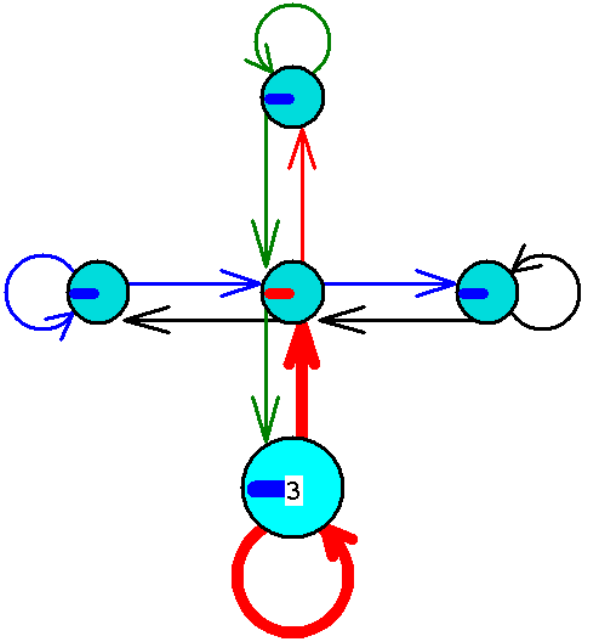
Голямата стъпка е изпълнението на алгоритъм. Това е основното предимство на новия език за описание на светове. В този нов език бъдещето не се предвижда, като се гледа няколко стъпки напред, а се гледа няколко големи стъпки напред.

По този начин може да се погледне по-далеч в бъдещето, защото, ако мислим в малки стъпки, ще се получи комбинаторна експлозия и няма да можем да видим по-далеч от носа си.

Алгоритъма на коня



Алгоритъма на топа



Последствия

В третата част на дисертацията се разглеждат някои философски въпроси, свързани с последствията от създаването на ИИ. Тази част не е математическа и вътре няма нито една теорема и дори нито една дефиниция. Все пак това е важна част от дисертацията, защото математиката не се състои само от построяването на формални конструкции и извършването на формални доказателства. Математикът е длъжен да зададе и въпроса: „Какви ще са последствията от тези формални разсъждения?“.

Трябва ли ИИ да бъде общодостъпна технологията?

Отговорът, който даваме, е, че това е опасна технология и че сериозните статии в областта на ИИ трябва да бъдат засекретени. Този отговор поставя в неудобно положение евентуалния бъдещ рецензент на дисертацията. Ако той приеме тази наша теза и ако смята, че тази дисертация е сериозен принос в областта на ИИ, то той трябва да се опита да я скрие от обществеността, като даде отрицателна рецензия. От друга страна, ако рецензентът приеме, че дисертацията не е сериозен принос в областта на ИИ, то той пак трябва да даде отрицателна рецензия.

Наукометрия

За последните 22 години съм публикувал 23 публикации по темата на дисертацията.

Цитатите по темата на дисертацията са 87,
от които 31 в Scopus, 5 в дисертации и 15 в книги.

Цитатите започнаха да се появяват 15 години след излизането на най-цитираната ми статия. За цитатите трябва да благодаря на Jose Hernandez-Orallo. Той пръв ме цитира и благодарение на това бях индексиран в Google и започнаха да ме цитират и други хора.

Критика

Дисертацията предизвика сериозна критика от страна на колегите за това, че изложението не е притежава нужната строгост и формалност. Искам да благодаря за конструктивната критика, благодарение на която подобрих изложението. Това най-добре се вижда в последната ми статия, където е дадена формалната дефиниция на ИИ.

Най-вече искам да благодаря за критиката на моя учител професор Димитър Скордев, когото наскоро загубихме. Той никога не е пестил забележките си към моята работа, но винаги критиката му е била добронамерена и градивна.

Професор Скордев беше ръководител на моята дипломна работа. Писането на тази дипломна работа продължи пет години. През тези пет години аз работех като демонстратор и водех упражненията към лекциите на проф. Скордев.

След като завърших СУ аз постъпих на работа в ИМИ и започнах да пиша докторската си дисертация. Научен ръководител ми беше доц. Иванов и проф. Скордев ми беше научен консултант. По-късно, когато смених темата на дисертацията, проф. Скордев отказа да ми бъде научен консултант като каза, че той не е компетентен да води дисертация по ИИ. Истината е, че той не вярваше в ИИ имаше предубеждение към тази тема. Много спорове сме водили по въпроса, но не успях да променя негативното му отношение към ИИ.

Въпреки всичко проф. Скордев продължи да следи работата ми и да ми помага без да пести критиката си. Последният път, когато ме критикува, беше това лято, когато му изпратих предпоследния вариант на тази дисертация. Тогава той ми изпрати сериозни забележки, част от които успях да отстраня в последния вариант на дисертацията.