



OŚRODEK
PRZETWARZANIA
INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Warsztaty z Algorytmu Optymalizacyjnego

Sebastian Fijałkowski

Maciej Kowalski

wrzesień 2021, Warszawa

Cel warsztatów



Przedstawienie postawionego problemu oraz jego ograniczeń.



Przedstawianie materiałów/ narzędzi dostarczonych przez OPI PIB.



Przedstawienie przykładowych rozwiązań postawionego problemu.



Wsparcie w przygotowaniu własnych rozwiązań przez zainteresowane podmioty.



Weryfikacja, czy posiadane przez podmioty rozwiązania są efektywne jakościowo (wynik punktowy) oraz wydajnościowo (czas realizacji).



Konsultacje dla podmiotów planujących przyjęcie rozwiązań informatycznych od podmiotów trzecich.

Cel optymalizacji dorobku naukowego

Celem działania algorytmu optymalizacyjnego jest zapewnienie podmiotom ewaluowanym najwyższego wyniku za osiągnięcia publikacyjne przy zadanych ograniczeniach (limitów instytucji, listy i limitu osób, listy i wartości udziałów).

Przestrzenią działania dla algorytmu optymalizacyjnego są wszystkie udziały w publikacjach naukowych.

Jego zadaniem jest znalezienie takiego zbioru udziałów, które będą spełniały:

- a) wszystkie ograniczenia wynikające z rozporządzenia w sprawie ewaluacji jakości działalności naukowej,
- b) suma punktów z tych udziałów da najwyższy możliwy wynik.

Skala problemu

Celem optymalizacji jest wskazanie takiej kombinacji udziałów, która daje najwyższy wynik oraz spełnia zadane ograniczenia.

Pragnąc sprawdzić wszystkie możliwe podzbiory udziałów, otrzymujemy olbrzymią liczbę kombinacji wielkości 2^N gdzie N to liczba wszystkich udziałów. Spotykamy się tu z wybuchem wykładniczym.

Przykładowo:

W największych podmioto-dyscyplinach mamy ponad 1000 autorów z liczbą udziałów pojedynczego autora nawet do 500. W podmioto-dyscyplinach z co najmniej kilkunastoma tysiącami udziałów, wykorzystanie siłowych metod (ang. brute force) jest niemożliwe do zastosowania ze względu na czasochłonność.

Co udostępniamy jako OPI PIB?

Ewaluator

narzędzie do wyliczenia punktowego wyniku i jego weryfikacji

Narzędzie w postaci aplikacji java, uruchamialne z linii poleceń z określonymi argumentami w postaci plików .csv.

W wyniku działania tego narzędzia dostajemy wynik punktowy oraz raport z weryfikacji ograniczeń ewaluacyjnych.

Materiały

Dokładny opis struktury i semantyki danych wejściowych.

Format danych wyjściowych – w dostarczonych danych testowych. W pliku udziałów w pozycji „ao_accepted” należy ustawiać odpowiednio true/false.

Dane testowe

Archiwum zip z danymi zanonimizowanymi dla 18 podmioto-dyscyplin o różnej liczbie udziałów.

Dane testowe

- Dane testowe są udostępnione na stronie warsztatów w postaci archiwum zip – external_tests.zip
- Archiwum zawiera 18 podmioto-dyscyplin, o zróżnicowanej liczbie udziałów od kilkuset do prawie 15 tys.
- Wszystkie dane uległy anonimizacji tzn. wszelkie oryginalne identyfikatory zostały zmienione na sztuczne.

id	institution_uuid	discipline_id	ilość udziałów
1	00000000-0000-0000-0000-000000000001	100	15340
2	00000000-0000-0000-0000-000000000002	100	11159
3	00000000-0000-0000-0000-000000000003	100	10311
4	00000000-0000-0000-0000-000000000004	100	6757
5	00000000-0000-0000-0000-000000000005	100	6089
6	00000000-0000-0000-0000-000000000006	100	4242
7	00000000-0000-0000-0000-000000000007	100	9929
8	00000000-0000-0000-0000-000000000008	100	3305
9	00000000-0000-0000-0000-000000000009	100	4686
10	00000000-0000-0000-0000-000000000010	100	3062
11	00000000-0000-0000-0000-000000000011	100	1158
12	00000000-0000-0000-0000-000000000012	100	1556
13	00000000-0000-0000-0000-000000000013	100	1720
14	00000000-0000-0000-0000-000000000014	100	790
15	00000000-0000-0000-0000-000000000015	100	284
16	00000000-0000-0000-0000-000000000016	100	1367
17	00000000-0000-0000-0000-000000000017	100	150
18	00000000-0000-0000-0000-000000000018	100	128

Dane testowe

Dane testowe są danymi wejściowymi dla procesu optymalizacji.

Dane zostały podzielone na trzy części – osobne dokumenty w formacie .csv:



Ograniczenia dla podmiotu



Ograniczenia dla osób



Udziały

We wszystkich dokumentach (format .csv) wprowadzono identyfikator instytucji oraz identyfikator dyscypliny – jest to para identyfikująca pojedynczą podmioto-dyscyplinę.

Dane testowe – ograniczenia dla podmiotu

```
institution_uuid;discipline_code;limit_3n_u;limit_opn;limit_mono_p;limit_m1921  
1a88b2ad-f6fd-4e5a-b514-27766a21594d;404;207;46.2;11.55;168.872
```

Dokument ograniczeń dla podmiotu zawiera:

- Ograniczenie 3N,
- Ograniczenie na liczbę udziałów osób spoza N,
- Ograniczenie na liczbę udziałów podmiotu,
- Ograniczenie na liczbę udziałów z lat 2019-2021.



Dane testowe – ograniczenia dla osób

```
institution_uuid;discipline_code;aux_person_id;person_type;limit_all_u;limit_mono_o  
1a88b2ad-f6fd-4e5a-b514-27766a21594d;404;2427>true;4;2  
1a88b2ad-f6fd-4e5a-b514-27766a21594d;404;2915>true;3.2;1.6  
1a88b2ad-f6fd-4e5a-b514-27766a21594d;404;3357>true;4;2  
(...)
```

Dokument ograniczeń osób zawiera:

- Identyfikator osoby,
- Informacje, czy osoba jest uwzględniona w liczbie N,
- Ograniczenie udziałów osoby ogółem,
- Ograniczenie udziałów osoby na monografii.



Dane testowe – udziały

```
institution_uuid;discipline_code;id;aux_person_id;person_type;mono_limit_p;mono_limit_o;m1921_limit;u;pu;source_id;source_limit;ao_accepted
1a88b2ad-f6fd-4e5a-b514-27766a21594d;404;396877;109338;true;false;false;false;0.3333;0.3333;;;false
1a88b2ad-f6fd-4e5a-b514-27766a21594d;404;630703;33441;true;false;true;true;0.25;1.25;;;false
1a88b2ad-f6fd-4e5a-b514-27766a21594d;404;269816;30597;true;false;false;false;0.5;0.5;;;false
(...)
```

Dokument z udziałami zawiera informacje:

- Identyfikator autora udziału,
- U – waga/wartość udziału,
- PU – punktacja za ten udział,
- Czy jest to udział osoby uwzględnionej w liczbie N,
- Czy udział zalicza się do limitu monografii osoby,
- Czy udział zalicza się do limitu monografii podmiotu,
- Czy udział zalicza się do limitu osiągnięć z lat 2019-2021 lub monografii,
- Opcjonalnie identyfikator monografii,
- Opcjonalnie limit punktów dla rozdziałów z monografii.



Ewaluator

Aplikacja Java (artefakt .jar), która jest uruchamiana z linii poleceń z argumentami wejściowymi w postaci plików .csv

Program przyjmuje trzy parametry wejściowe:

- plik instytucji (w formacie .csv), zawierający jeden rekord opisujący podmioto-dyscyplinę,
- plik osób (w formacie .csv) zawierający wiele rekordów opisujących właścicieli udziałów oraz ich limity,
- plik udziałów (w formacie .csv) zawierający wiele rekordów opisujących udziały wraz z ich wartościami.

Ewaluator – przykład wywołania

```
D:\evaluator>java -jar sedn-algo-evaluator-1.0.jar ./inst.csv ./persons.csv ./shares.csv
13:51:00.452 [main] INFO pl.org.opi.sedn.algo.service.AchievementManager - Total points: 352,057
13:51:00.455 [main] INFO pl.org.opi.sedn.algo.service.AchievementManager - Score (not validated): 7341,6094
13:51:00.456 [main] INFO pl.org.opi.sedn.algo.service.AchievementManager - Opn value: 0, Limit: 46,2
13:51:00.456 [main] INFO pl.org.opi.sedn.algo.service.AchievementManager - U value: 206,9891, Limit: 207
13:51:00.457 [main] INFO pl.org.opi.sedn.algo.service.AchievementManager - M1921 value: 112,383, Limit: 168,872
13:51:00.457 [main] INFO pl.org.opi.sedn.algo.service.AchievementManager - MonoP value: 1, Limit: 11,55
13:51:00.457 [main] INFO pl.org.opi.sedn.algo.service.AchievementManager - ---
13:51:00.457 [main] INFO pl.org.opi.sedn.algo.service.AchievementManager - People over limit:
13:51:00.458 [main] INFO pl.org.opi.sedn.algo.service.AchievementManager - ---
13:51:00.458 [main] INFO pl.org.opi.sedn.algo.service.AchievementManager - Sources over limit:
13:51:00.459 [main] INFO pl.org.opi.sedn.algo.service.AchievementManager - ---
13:51:00.460 [main] INFO pl.org.opi.sedn.algo.service.AchievementManager - Score (validated): 7341,6094
```

Problem optymalizacyjny

- Udziały, Osoby, Instytucja (podmioto-dyscypliny)

	institution_uuid	discipline_code	id	aux_person	person_type	mono_limit_r	mono_limit_c	m1921_limit	u	pu	source_id	source_limit	ao_accepted
	integer	integer	bigint	bigint	boolean	boolean	boolean	boolean	numeric	numeric	text	integer	boolean
1	a013df22-d3	203	391628	128861	TRUE	FALSE	FALSE	TRUE	0.5774	40.4180			FALSE
2	fe0251f2-37	203	27609	169749	TRUE	FALSE	FALSE	TRUE	0.7071	49.4970			FALSE
3	4e6a4738-94	506	67398	41727	TRUE	FALSE	FALSE	TRUE	0.7071	49.4970			FALSE
4	b8378660-15	503	343937	93220	TRUE	FALSE	FALSE	TRUE	0.5000	35.0000			FALSE
5	b8378660-15	503	343936	166248	TRUE	FALSE	FALSE	TRUE	0.5000	35.0000			FALSE
6	4e6a4738-94	506	461884	124790	TRUE	FALSE	FALSE	TRUE	1.0000	70.0000			FALSE
7	f8d0b933-97	506	179302	55342	TRUE	FALSE	FALSE	TRUE	0.7071	49.4970			FALSE
8	4e6a4738-94	506	137573	107135	TRUE	FALSE	FALSE	TRUE	1.0000	70.0000			FALSE
9	4e6a4738-94	506	16861	22854	TRUE	FALSE	FALSE	TRUE	1.0000	70.0000			FALSE
10	4e6a4738-94	506	523057	16366	TRUE	FALSE	FALSE	TRUE	1.0000	70.0000			FALSE
11	29483405-4e	503	482170	93072	TRUE	FALSE	FALSE	TRUE	0.7071	49.4970			FALSE
12	f8d0b933-97	506	361457	47560	TRUE	FALSE	FALSE	TRUE	0.5774	40.4180			FALSE
13	29483405-4e	503	119555	140701	TRUE	FALSE	FALSE	TRUE	1.0000	70.0000			FALSE
14	1c5fb903-cc	506	118372	52307	TRUE	FALSE	FALSE	TRUE	1.0000	70.0000			FALSE
15	a013df22-d5	605	265988	89091	TRUE	FALSE	FALSE	TRUE	1.0000	140.0000			FALSE
16	dbdda9a7-f6	203	168791	104577	TRUE	FALSE	FALSE	TRUE	1.0000	140.0000			FALSE
17	5011bcd5-df	501	495231	50027	TRUE	FALSE	FALSE	FALSE	1.0000	15.0000			FALSE
18	b8378660-15	206	591973	168659	TRUE	FALSE	FALSE	FALSE	0.1667	2.5005			FALSE

- Szukamy wektora wynikowego (wektora, którego elementy określają, czy dany udział wchodzi w skład ewaluowanego „plecaka”)

X =	1	0	1	1	0	1	0	0	...	0
-----	---	---	---	---	---	---	---	---	-----	---

Wszystko byłoby proste...

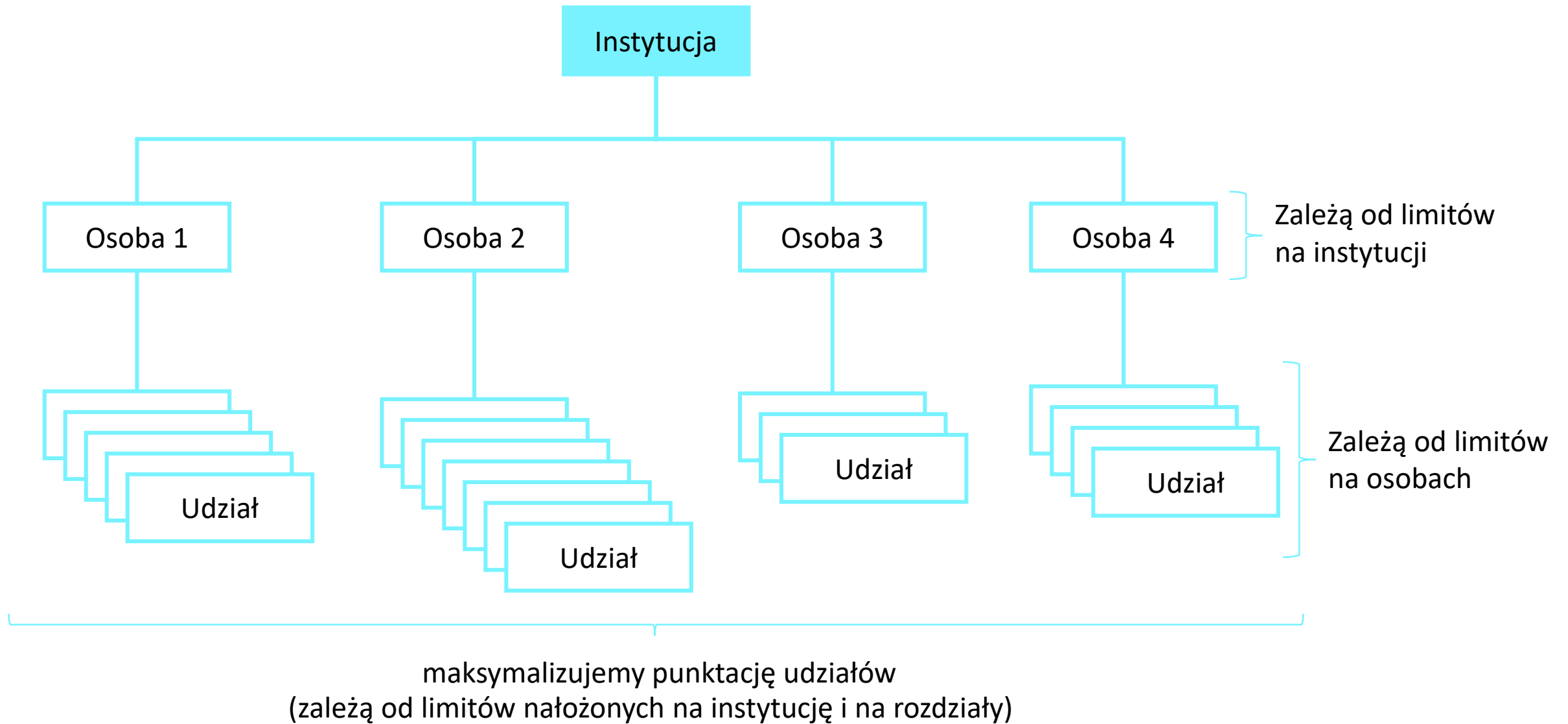
- Maksymalizujemy sumę wartości udziałów

x =	1	0	1	1	0	1	0	0	...	0
u =	1	0.5	0.5	0.25	0.333	1	0.5	0.5	...	1
pu =	200	50	1	12	12	175	70	20	...	10

$$\operatorname{argmax}(x \bullet pu')$$

- Ale są limity:
 - Limit udziałów osób spoza N,
 - Limit udziałów osób z 3N,
 - Limit udziałów z publikacji monografii podmiotów za lata 2019-2021,
 - Limit udziałów monografii podmiotów,
 - Limit udziałów monografii osoby,
 - Limit rozdziałów (suma udziałów rozdziałów nie może być większa niż udział monografii).

Ilustracja problemu



Constraint programming

- Mamy wektory $x = [x_1, x_2, x_3, x_4]$, $u = [u_1, u_2, u_3, u_4]$, $pu = [pu_1, pu_2, pu_3, pu_4]$
- Chcemy ustalić wartości x w taki sposób, żeby:

Limity na udziały instytucji

$$\begin{cases} u_1x_1 + u_2x_2 + u_3x_3 + u_4x_4 \leq limit_3n \\ u_1x_1 + u_2x_2 + u_3x_3 + u_4x_4 \leq limit_m1921 \\ \dots \end{cases}$$

Limity na udziały dla osób

$$\begin{cases} 0x_1 + u_2x_2 + u_3x_3 + 0x_4 \leq limit_osoby_1 \\ u_1x_1 + 0x_2 + 0x_3 + u_3x_3 \leq limit_osoby_1 \\ \dots \end{cases}$$

Limity na udziały dla rozdziałów:

$$\begin{cases} 0x_1 + u_2x_2 + u_3x_3 + 0x_4 \leq limit_rozdziółow_1 \\ u_1x_1 + 0x_2 + 0x_3 + 0x_3 \leq limit_rozdziółow_2 \\ \dots \end{cases}$$

Constraint programming (2)

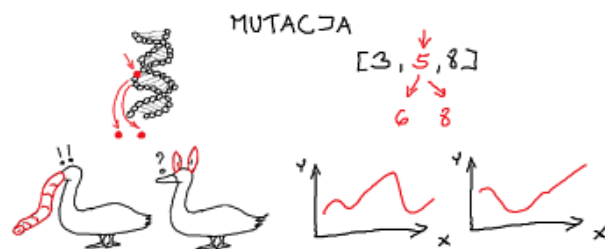
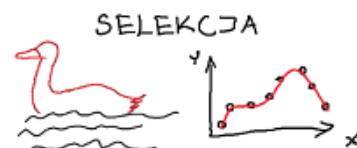
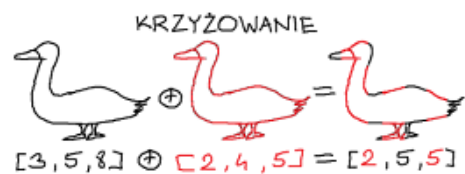
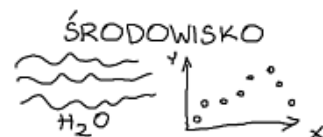
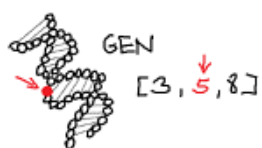
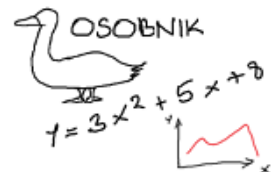
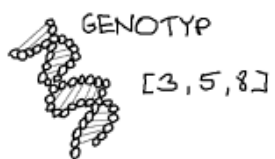
- Oraz żeby zmaksymalizować wartość:

$$\operatorname{argmax}(pu_1x_1 + pu_2x_2 + pu_3x_3 + pu_4x_4)$$

Algorytm genetyczny

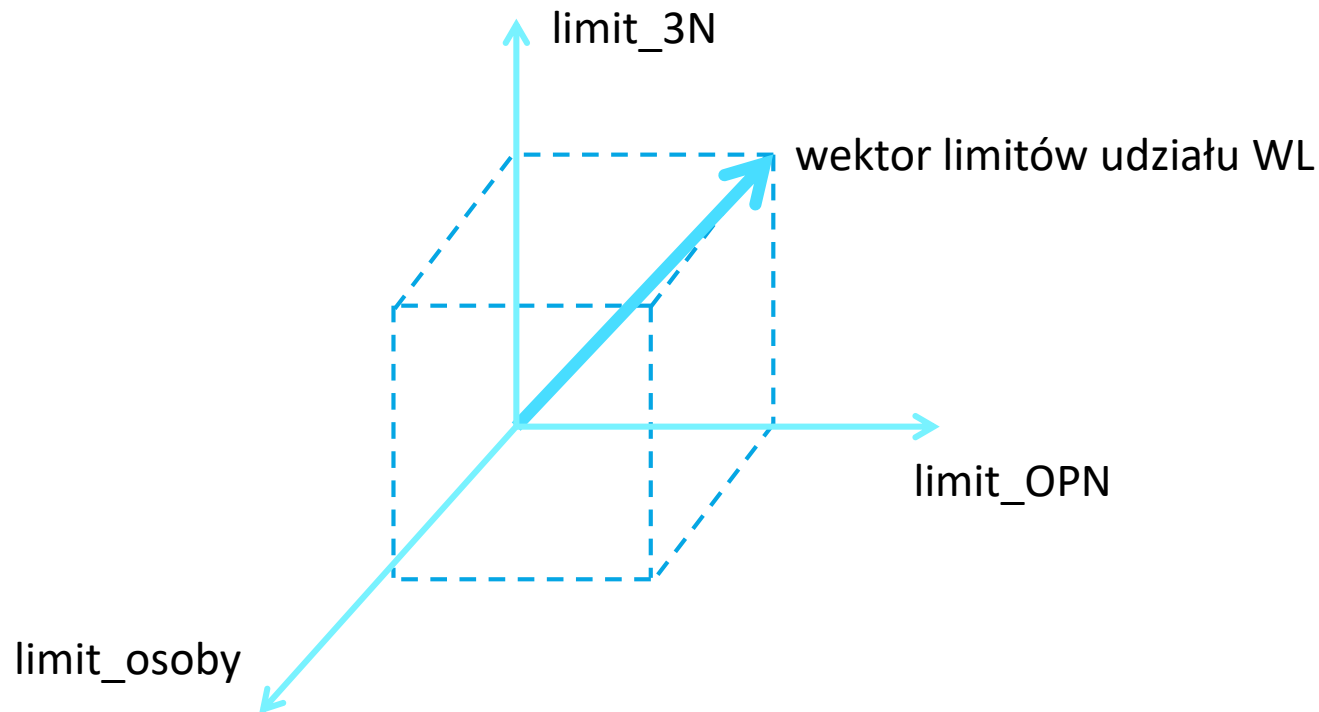
- Sprytnie wyznaczamy sobie wektory początkowe
- Dla każdego sprawdzamy jaki daje wynik (zastosowanie tego wektora do limitu, jakie daje to punkty). Jeżeli któryś z limitów jest przekroczony, wtedy wynikiem takiej optymalizacji jest wartość „0”.
- Krzyżujemy, mutujemy, a powstałe wektory tworzą nową „generację” (epokę).
- Zastosowanie funkcji celu (im większy wynik, tym lepiej dopasowany wektor), umożliwia coraz lepsze dopasowanie wektora w kolejnych generacjach.
- Proces jest zatrzymywany, jeżeli algorytm osiągnął stabilność.

ALGORYTM GENETYCZNY



Co to znaczy sprytnie? (1)

- To ile limitów zabiera dany udział, możemy traktować jako wektor (przestrzenią tego wektora będą poszczególne limity).



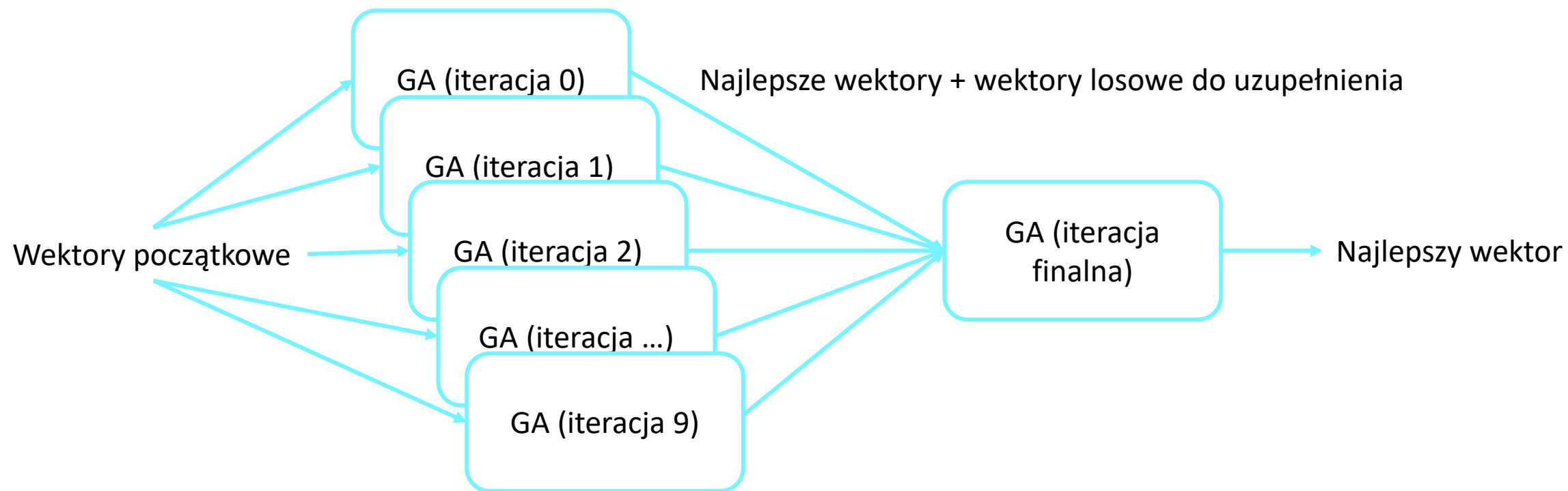
Co to znaczy sprytnie? (2)

1. Zależy nam na wektorach, które zabierają najmniej z limitów, przy jednoczesnej dużej wartości punktów.

- Sortujemy udziały według kryterium: $pu/|WL|$,
- Iterujemy:
 - Jeżeli wzięcie tego udziału przekroczy nam limity, to go odrzucamy;
 - Jeżeli wzięcie tego udziału nie przekracza limitów, to go bierzemy;
 - Bierzemy następny udział.

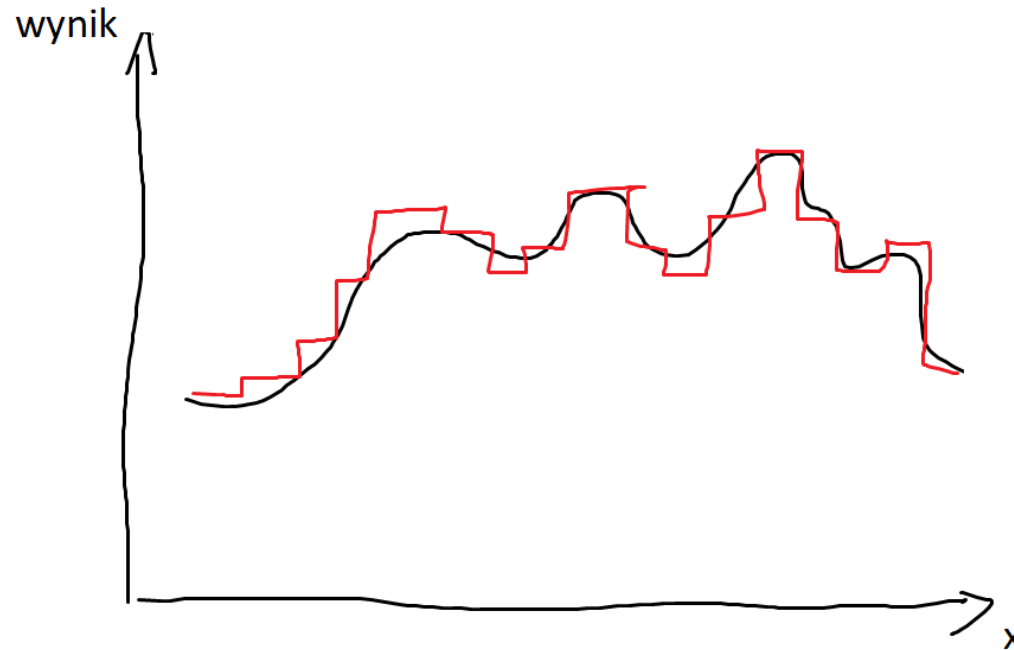
2. Wykorzystanie dostępnych solverów np. ORTools, dzięki którym otrzymujemy bazę do algorytmu genetycznego.

Etap GA



Założenia

- Przeszukujemy pewnego rodzaju funkcję rozpiętą w hiperprzestrzeni, która posiada minima i maksima.
- Funkcja, którą przeszukujemy stabilna (?)



Ograniczenia

- Możliwość wpadnięcia w lokalne maksimum (ograniczamy taką możliwość poprzez stosowanie kaskad GA). Sam GA dzięki zastosowaniu mutacji (oraz dużej ilości czasu) może „wyskoczyć” z takiego lokalnego maksimum.
- GA nie gwarantuje najlepszego możliwego wyniku. Wraz ze wzrostem ilości epok możemy jedynie uprawdopodobniać wynik.

Wyniki modelu OPI PIB na danych testowych

id	institution_uuid	discipline_id	ilość udziałów	wynik OPI
1	00000000-0000-0000-0000-000000000001	100	15340	163862,2209
2	00000000-0000-0000-0000-000000000002	100	11159	132823,3247
3	00000000-0000-0000-0000-000000000003	100	10311	82607,7284
4	00000000-0000-0000-0000-000000000004	100	6757	58079,3597
5	00000000-0000-0000-0000-000000000005	100	6089	59780,5467
6	00000000-0000-0000-0000-000000000006	100	4242	36678,7625
7	00000000-0000-0000-0000-000000000007	100	9929	62827,7434
8	00000000-0000-0000-0000-000000000008	100	3305	35265,4492
9	00000000-0000-0000-0000-000000000009	100	4686	67582,041
10	00000000-0000-0000-0000-000000000010	100	3062	27395,7969
11	00000000-0000-0000-0000-000000000011	100	1158	23988,9329
12	00000000-0000-0000-0000-000000000012	100	1556	18723,3659
13	00000000-0000-0000-0000-000000000013	100	1720	16278,689
14	00000000-0000-0000-0000-000000000014	100	790	11727,695
15	00000000-0000-0000-0000-000000000015	100	284	6019,3292
16	00000000-0000-0000-0000-000000000016	100	1367	8330,6437
17	00000000-0000-0000-0000-000000000017	100	150	1583,5
18	00000000-0000-0000-0000-000000000018	100	128	4414,0475

Co dalej

1

Po pierwszym spotkaniu:

Uczestnicy rejestrują się na stronie warsztatów, co umożliwia pobranie m.in. Ewaluatora.

Uczestnicy rozpoczynają eksperymentowanie z danymi oraz budowanie swoich rozwiązań.

Uczestnicy zobowiązują się do późniejszego przedstawienia swojego podejścia do problemu, metodyki rozwiązania problemu, czy doświadczeń z jego implementacji, w tym ich wyników.

2

Drugie spotkanie:

W trakcie prac nad swoimi rozwiązaniami uczestnicy mogą mieć wiele pytań, obserwacji, które warto omówić.

Jest to spotkanie techniczne mające na celu wymianę dotychczasowych doświadczeń, spostrzeżeń oraz rozwiązanie wątpliwości.

3

Trzecie spotkanie:

Spotkanie podsumowujące warsztaty, pole na podzielenie się uzyskanymi wynikami przez różnych uczestników.

Każdy z uczestników przedstawi swoje podejście, jego wyniki na danych testowych, jego zalety i wady.

Każdy z uczestników szczegółowo omówi stabilności wyników między wywołaniami, czy czas potrzebny na przeliczenie danych testowych.



OŚRODEK
PRZETWARZANIA
INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Dziękujemy za uwagę

Ośrodek Przetwarzania Informacji
Państwowy Instytut Badawczy
al. Niepodległości 188 B
00-608 Warszawa

tel.: +48 22 570 14 00
faks: +48 22 825 33 19
e-mail: opi@opi.org.pl
www.opi.org.pl