Informatyka, studia STACJONARNE, I st. semestr VI

Komputerowe systemy rozpoznawania **2020/2021**

Prowadzący: prof. dr hab. inż. Adam Niewiadomski poniedziałek, 12:00

Data oddania: _____ Ocena: _____

Hubert Gawłowski 224298
Kamil Kiszko-Zgierski 224328

Projekt 2. Podsumowania lingwistyczne relacyjnych baz danych

Opis projektu ma formę artykułu naukowego lub raportu z zadania badawczego/doświadczalnego/obliczeniowego (wg indywidualnych potrzeb związanych np. z pracą inżynierską/naukową/zawodową).

Wybrane sekcje (rozdziały sprawozdania) są uzupełniane wg wymagań w opisie Projektu 2. i Harmonogramie Zajęć na WI-KAMP KSR jako efekty zadań w poszczególnych tygodniach.

1. Cel

Celem zadania jest zaimplementowanie lingwistycznej agregacji, tj. przedstawienie danych liczbowych za pomocą sformułowań w języku pozornie naturalnym. Owa implementacja zostanie wykonana w technologii Java z graficznym interfejsem użytkownika oraz z wykorzystaniem systemu zarządania bazą danych o nazwie MySql, a badania zostaną przeprowadzone w oparciu o bazę danych zawierającą statystyki zawodników ligi koszykówki NBA od sezonu 1996/1997 do 2019/2020. [4].

2. Charakterystyka podsumowywanej bazy danych

Jako bazę danych w zadaniu wykorzystaliśmy bazę danych dotyczącą informacji o zawodnikach NBA (najbrdziej prestiżowej lidze koszykarskiej na świecie) z lat 1996 - 2019. Cała baza danych składa się z 22 kolumn zawierających informacje o danych zawodnika oraz jego statystykach różnego rodzaju.

Baza danych ma charakter pliku csv i zawiera 11144 rekordy. Spośród kolumn wybraliśmy te, które zawierają wartości, naszym zdaniem, najlepiej nadające się do rozmycia. Atrybuty do rozmycia dobraliśmy tak, aby zgadzały się z definicją zmiennej lingwistycznej, która prezentuje się następująco [1]:

$$L =< \mathcal{L}, H(L), X, G, K > \tag{1}$$

gdzie:
$\mathcal{L}$ jest nazwą zmiennej lingwistycznej L,
H lub H(L) jest zbiorem terminów lingwistycznych $l_1, l_2, ..., l_N, N \in \mathbf{N}$, które są wartościami L,
X jest przestrzenią rozważań, w której określa się zbiory rozmyte $S_1, S_2, ..., S_N, N \in \mathbf{N}$, reprezentujące odpowiednio terminy $l_1, l_2, ..., l_N$,
G jest regułą gramatyczną, która generuje terminy w zbiorze $H(L)$,
K jest regułą semantyczną, która przyporządkowuje terminom l_i zbiory rozmyte S_i 2 X, czyli $K : l_i \rightarrow S_i, S_i = \{< x, \mu_{S_i}(x) >: x \in X\}, i = 1, 2, ..., N$

Ostatecznie spośród 22 kolumn wybraliśmy następujące:

— Kolumny nie dające się rozmyć (zawierające informacje pozwalające zidentyfikować zawodnika):
1. nr identyfikacyjny zawodnika (index)
2. imię i nazwisko zawodnika (player_name)
3. skrót nazwy drużyny (team_abbreviation)
4. kraj urodzenia zawodnika (country)

— Kolumny dające się rozmyć:
1. wiek zawodnika (age) - atrybut przyjmuje wartości liczbowe całkowite od 18 do 44. Przekładając ten atrybut na język naturalny możemy użyć wartości: "junior", "młody", "w średnim wieku", "doświadczony", "stary".
2. wzrost zawodnika (player_height) - atrybut przyjmuje wartości liczbowe zmiennoprzecinkowe od 160,02 do 231,14. Przekładając ten atrybut na język naturalny możemy użyć wartości: "bardzo niski", "niski", "średniego wzrostu", "wysoki", "bardzo wysoki".
3. waga zawodnika (player_weight) - atrybut przyjmuje wartości liczbowe zmiennoprzecinkowe od 60,33 do 163,29. Przekładając ten atrybut na język naturalny możemy użyć wartości: "bardzo lekki", "lekki", "o przeciętnej wadze", "ciężki", "bardzo ciężki".
4. numer w drafcie (draft_number) - oznacza, który w kolejności został wybrany zawodnik w organizowanym co roku tzw. drafcie, który polega na wybieraniu przez drużyny młodych zawodników na następny sezon. Atrybut przyjmuje wartości liczbowe całkowite od 1 do 60 oraz wartość "undrafted" - nie wybrany w drafcie. Wartość "undrafted będzie traktowana jako wartość maksymalna. Przekładając ten atrybut na język naturalny możemy użyć wartości: "Bardzo szybko", "szybko", "średnio", "późno", "bardzo późno", "nie wybrano".
5. rozegrane gry (gp) - liczba gier zawodnika w danym sezonie. Atrybut ten przyjmuje wartości liczbowe całkowite od 1 do 85. Przekładając ten atrybut na język naturalny możemy użyć wartości: "bardzo mało", "mało", "średnio", "dużo", "bardzo dużo".

6. zdobyte punkty (pts) - średnia liczba punktów zdobytych przez zawodnika w meczach w danym sezonie. Atrybut ten przyjmuje wartości liczbowe zmiennoprzecinkowe od 0 do 36,1. Przekładając ten atrybut na język naturalny możemy użyć wartości: "niewiarygodnie mało", "mało", "dostatecznie", "dużo", "nieziemsko dużo".
7. liczba zbiórek (reb) - średnia liczba zbiórek zawodnika na mecz - zbiórka jest to złapanie piłki przez zawodnika po nieudanym rzucie do kosza. Atrybut ten przyjmuje wartości liczbowe zmiennoprzecinkowe od 0 do 16,3. Przekładając ten atrybut na język naturalny możemy użyć wartości: "bardzo mało", "mało", "dostatecznie", "dużo", "bardzo dużo".
8. liczba asyst (ast) - średnia liczba asyst zawodnika na mecz - asysta jest to ostatnie podanie między zawodnikami tej samej drużyny, po którym zdobyty zostaje punkt. Atrybut ten przyjmuje wartości liczbowe zmiennoprzecinkowe od 0 do 11,7. Podobnie jak w atrybucie wyżej, przekładając ten atrybut na język naturalny możemy użyć wartości: "bardzo mało", "mało", "dostatecznie", "dużo", "bardzo dużo".
9. wpływ na drużynę (net_rating) - jaki wpływ miał zawodnik na punkty drużyny, gdy znajdował się na parkiecie. Atrybut ten przyjmuje wartości liczbowe zmiennoprzecinkowe od -100,0 do 100,0 ze zdecydowaną większością wartości mieszczących się w przedziale $< -25,0; 25,0 >$. Przekładając ten atrybut na język naturalny możemy użyć wartości: "bardzo negatywny", "negatywny", "neutralny", "pozytywny", "bardzo pozytywny"
10. skuteczność rzutów (ts_pct) - jak efektywnie zawodnik rzucał do kosza, czyli ile spośród jego rzutów kończyło się punktami. Atrybut ten przyjmuje wartości liczbowe zmiennoprzecinkowe od 0 do 1, gdzie 0 oznacza, że żaden rzut nie kończył się punktem, a 1 - każdy rzut zawodnika kończył się punktem. Najwięcej wartości znajduje się w przedziale: $< 0,30; 0,67 >$. Przekładając ten atrybut na język naturalny możemy użyć wartości: "dramatycznie nieskuteczny", "nieskuteczny", "przeciętny", "skuteczny", "niesamowicie skuteczny".
11. procent asyst (ast_pct) - procent punktów przy jakich asystował zawodnik, gdy znajdował się na boisku - atrybut ten mówi dużo o wpływie zawodnika na drużynę. Przyjmuje on wartości liczbowe zmiennoprzecinkowe z przedziału $< 0; 1 >$ ze zdecydowaną większością wartości w przedziale $< 0; 0,40 >$. Przekładając ten atrybut na język naturalny możemy użyć wartości: "bardzo mały", "mały", "przeciętny", "duży", "bardzo duży".

Oczywiście wartości takie jak: wzrost, waga, czy wiek będą odnosiły się jedynie do koszykarzy, ponieważ np. słowo "stary" w kontekście gracza koszykówki będzie związane z zupełnie innym wiekiem, niż słowo "stary" używane na codzień w stosunku do opisu wieku ludzi.

W rankingu najbardziej dochodowych lig sportowych liga NBA zajmuje

trzecie miejsce na świecie [1] (wyżej w rankingu są tylko ligi NFL - hokej i MLB - baseball). W związku z dużym zainteresowaniem wokół niej zasadnym jest, aby dane liczbowe koszykarzy przedstawiać również za pomocą zmiennych lingwistycznych.

Po pierwsze, jako iż owa dyscyplina jest bardzo popularna, wielu kibiców jest również zainteresowanych statystykami koszykarzy. Dla przeciętnego widza dane liczbowe mogą być jednak mało zrozumiałe. Rozwiązaniem tego problemu wydaje się wprowadzenie zmiennych lingwistycznych w celu ułatwienia interpretacji danych, co wiąże się z lepszym przyswojeniem informacji przez odbiorcę, jak również zaoszczędzeniem czasu na próbie ich zrozumienia.

Po drugie, przedstawienie danych liczbowych w postaci zmiennych lingwistycznych zapewnia grupowanie rekordów. Dzięki temu, dużo łatwiejsze jest dokonanie filtracji zawodników, co skutkuje szybszym wyszukaniem graczy o podobnych parametrach fizycznych lub osiąganych statystykach.

W końcu, popularność oraz statystyki przychodów klubów NBA powodują zainteresowanie wśród potencjalnych sponsorów, którzy niekoniecznie muszą być związani z koszykówką. Dane przedstawione w formie zmiennych lingwistycznych znacznie pomogłyby w analizie statystyk oraz ocenie ryzyka związanego z inwestycją w dany klub.

#	Field	Schema	Table	Type
1	index	nba_players	nba_players	INT
2	player_name	nba_players	nba_players	VARCHAR
3	team_abbreviation	nba_players	nba_players	VARCHAR
4	country	nba_players	nba_players	VARCHAR
5	age	nba_players	nba_players	INT
6	player_height	nba_players	nba_players	DOUBLE
7	player_weight	nba_players	nba_players	DOUBLE
8	draft_number	nba_players	nba_players	VARCHAR
9	gp	nba_players	nba_players	INT
10	pts	nba_players	nba_players	DOUBLE
11	reb	nba_players	nba_players	DOUBLE
12	ast	nba_players	nba_players	DOUBLE
13	net_raiting	nba_players	nba_players	DOUBLE
14	ts_pct	nba_players	nba_players	DOUBLE
15	ast_pct	nba_players	nba_players	DOUBLE

Rysunek 1. Nazwy kolumn (atrybutów) w sytemie zarządzania bazami danych MySQL

3. Atrybuty i liczności obiektów wyrażone zmiennymi lingwistycznymi

Zmienne lingwistyczne dla wybranych 10 atrybutów z bazy danych, przedstawione w formie wykresów funkcji przynależności i wzorów analitycznych, wymienione etykiety oraz objaśnione wszystkie symbole ułatwiające czytelnikowi ich zrozumienie [2]. Zbędne jest cytowanie definicji. Konieczne precyzyjnie podane przestrzenie rozważań każdej zmiennej lingwistycznej, wzory i

[1] https://globalsportmatters.com/business/2019/03/07/tv-is-biggest-driver-in-global-sport-league-revenue/

wykresy dla każdej wartości/etykiety.
Jw. kwantyfikatory lingwistyczne – opisane etykietami, wykresami funkcji przynależności i wzorami analitycznymi. Uzasadnione wiedzą dziedzinową wybrane zakresy i etykiety. Precyzyjnie podane przestrzenie rozważań każdego kwantyfikatora lingwistycznego/rozmytego, wzory i wykresy dla każdej wartości/etykiety. Opisy własne z przypisami do literatury, tak by inżynier innej specjalności zrozumiał dalszy opis tego konkretnego ćwiczenia/eksperymentu.
Sekcja uzupełniona jako efekt zadania Tydzień 09 wg Harmonogramu Zajęć na WIKAMP KSR.

4. Narzędzia obliczeniowe: projekt (wybór, implementacja) i diagram UML pakietu obliczeń rozmytych. Diagram UML generatora podsumowań

4.1. Diagram pakietu obliczeń rozmytych

Diagram UML i zwięzły opis pakietu obliczeń rozmytych: źródło pakietu (zewnętrzny/własny/hybrydowy), przypis do literatury. Krótka charakterystyka najważniejszych klas i podstawowych dla zadania ich metod.
Sekcja uzupełniona jako efekt zadania Tydzień 10 wg Harmonogramu Zajęć na WIKAMP KSR.

4.2. Diagram UML generatora podsumowań. Krótka instrukcja użytkownika

Diagram UML generatora podsumowań (warstwy obliczeniowej oraz interfejsu użytkownika). Krótki ilustrowany opis jak użytkownik może korzystać z aplikacji, w szczególności wprowadzać parametry podsumowań, odczytywać wyniki oraz definiować własne etykiety i kwantyfikatory. Wersja JRE i inne wymogi niezbędne do uruchomienia aplikacji przez użytkownika na własnym komputerze.
Sekcja uzupełniona jako efekt zadania Tydzień 11 wg Harmonogramu Zajęć na WIKAMP KSR.

5. Jednopodmiotowe podsumowania lingwistyczne. Miary jakości, podsumowanie optymalne

Wyniki kolejnych eksperymentów wg punktów 2.-4. opisu projektu 2. Listy podsumowań jednopodmiotowych i tabele/rankingi podsumowań dla danych atrybutów obowiązkowe i dokładnie opisane w „captions” (tytułach), konieczny opis kolumn i wierszy tabel. Dla każdego podsumowania podane miary jakości oraz miara jakości podsumowania optymalnego.
Sekcja uzupełniona jako efekt zadania Tydzień 11 wg Harmonogramu Zajęć na WIKAMP KSR.

6. Wielopodmiotowe podsumowania lingwistyczne i ich miary jakości

Wyniki kolejnych eksperymentów wg punktów 2.-4. opisu projektu 2. Uzasadnienie i metoda podziału zbioru danych na rozłączne podmioty. Listy podsumowań wielopodmiotowych i tabele/rankingi podsumowań dla danych atrybutów obowiązkowe i dokładnie opisane w „captions" (tytułach), konieczny opis kolumn i wierszy tabel. Konieczne uwzględnienie wszystkich 4-ch form podsumowań wielopodmiotowych.

** Możliwe sformułowanie zagadnienia wielopodmiotowego podsumowania optymalnego **.

Ewentualne wyniki realizacji punktu „na ocenę 5.0" wg opisu Projektu 2. i ich porównanie do wyników z części obowiązkowej.

Sekcja uzupełniona jako efekt zadania Tydzień 12 wg Harmonogramu Zajęć na WIKAMP KSR.

7. Dyskusja, wnioski

Dokładne interpretacje uzyskanych wyników w zależności od parametrów klasyfikacji opisanych w punktach 3.-4 opisu Projektu 2. Szczególnie istotne są wnioski o charakterze uniwersalnym, istotne dla podobnych zadań. Omówić i wyjaśnić napotkane problemy (jeśli były). Każdy wniosek/problem powinien mieć poparcie w przeprowadzonych eksperymentach (odwołania do konkretnych wyników: tabel i miar jakości). Ocena które wybrane kwantyfikatory, sumaryzatory, kwalifikatory i/lub ich miary jakości mają małe albo duże znaczenie dla wiarygodności i jakości otrzymanych agregacji/podsumowań. Dla końcowej oceny jest to najważniejsza sekcja sprawozdania, gdyż prezentuje poziom zrozumienia rozwiązywanego problemu.

** Możliwości kontynuacji prac w obszarze logiki rozmytej i wnioskowania rozmytego, zwłaszcza w kontekście pracy inżynierskiej, magisterskiej, naukowej, itp. **

Sekcja uzupełniona jako efekt zadań Tydzień 11 i Tydzień 12 wg Harmonogramu Zajęć na WIKAMP KSR.

8. Braki w realizacji projektu 2.

Wymienić wg opisu Projektu 2. wszystkie niezrealizowane obowiązkowe elementy projektu, ewentualnie podać merytoryczne (ale nie czasowe) przyczyny tych braków.

Literatura

[1] A. Niewiadomski, Zbiory rozmyte typu 2. Zastosowania w reprezentowaniu informacji. Seria „Problemy współczesnej informatyki" pod redakcją L. Rutkowskiego. Akademicka Oficyna Wydawnicza EXIT, Warszawa, 2019.

[2] S. Zadrożny, Zapytania nieprecyzyjne i lingwistyczne podsumowania baz danych, EXIT, 2006, Warszawa

[3] A. Niewiadomski, Methods for the Linguistic Summarization of Data: Applications of Fuzzy Sets and Their Extensions, Akademicka Oficyna Wydawnicza EXIT, Warszawa, 2008.

[4] Baza danych zawierająca statystyki koszykarzy z ligi NBA z lat 1996-2020 [przeglądany 04.05.2021] Dostępna w: https://www.kaggle.com/justinas/nba-players-data

Literatura zawiera wyłącznie źródła recenzowane i/lub o potwierdzonej wiarygodności, możliwe do weryfikacji i cytowane w sprawozdaniu.