AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE
WYDZIAŁ ELEKTROTECHNIKI, AUTOMATYKI, INFORMATYKI I INŻYNIERII BIOMEDYCZNEJ

KATEDRA AUTOMATYKI I INŻYNIERII BIOMEDYCZNEJ

Praca dyplomowa magisterska

System szybkiego inteligentnego asocjacyjnego wyszukiwania relacji pomiędzy danymi wykorzystujący asocjacyjne grafowe struktury danych AGDS.

System of fast intelligent associative search for relations between data using associative graph data structures AGDS.

Autor: *Paulina Sopata*
Kierunek studiów: Automatyka i Robotyka
Opiekun pracy: *dr hab. Adrian Horzyk*

Kraków, 2017

Uprzedzony o odpowiedzialności karnej na podstawie art. 115 ust. 1 i 2 ustawy z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych (t.j. Dz.U. z 2006 r. Nr 90, poz. 631 z późn. zm.): „ Kto przywłaszcza sobie autorstwo albo wprowadza w błąd co do autorstwa całości lub części cudzego utworu albo artystycznego wykonania, podlega grzywnie, karze ograniczenia wolności albo pozbawienia wolności do lat 3. Tej samej karze podlega, kto rozpowszechnia bez podania nazwiska lub pseudonimu twórcy cudzy utwór w wersji oryginalnej albo w postaci opracowania, artystyczne wykonanie albo publicznie zniekształca taki utwór, artystyczne wykonanie, fonogram, wideogram lub nadanie.", a także uprzedzony o odpowiedzialności dyscyplinarnej na podstawie art. 211 ust. 1 ustawy z dnia 27 lipca 2005 r. Prawo o szkolnictwie wyższym (t.j. Dz. U. z 2012 r. poz. 572, z późn. zm.) „Za naruszenie przepisów obowiązujących w uczelni oraz za czyny uchybiające godności studenta student ponosi odpowiedzialność dyscyplinarną przed komisją dyscyplinarną albo przed sądem koleżeńskim samorządu studenckiego, zwanym dalej „sądem koleżeńskim", oświadczam, że niniejszą pracę dyplomową wykonałam osobiście i samodzielnie i że nie korzystałam ze źródeł innych niż wymienione w pracy.

Spis treści

1.Wstęp

W dzisiejszych czasach nie sposób wyobrazić sobie prawidłowego funkcjonowania potężnych instytucji, takich jak banki, szpitale, czy też urzędy, bez możliwości gromadzenia, przechowywania oraz wykorzystywania ogromnej ilości danych. Rola, jaką pełni informacja w funkcjonowaniu dzisiejszego świata jest niezwykle istotna, a obecne zdobycze technologii dają ogromne możliwości w formułowania złożonych wniosków na podstawie dogłębnej analizy relacji między informacjami.

Ludzki umysł, pomimo tego, iż jest potężnym narzędziem, dzięki któremu mamy dostęp do wielu algorytmów, metod oraz technik, bez których trudno wyobrazić sobie postęp informatyki, ma jednak dużo ograniczeń. Najważniejszymi z nich są ograniczenia pamięciowe, percepcyjne oraz skojarzeniowe. Obecnie bardzo istotnym jest poradzić sobie z trudnościami wynikającymi z możliwością szybkiego wyszukiwania oraz efektywnego analizowania ogromnych ilości informacji, gdzie należy wziąć pod uwagę wiele kombinacji, permutacji czy wariacji. Powiązanie danych ma na celu wyprowadzenie złożonych wniosków, czy przedstawienie jak najbardziej optymalnych wyników. Oprócz problemu z czasem potrzebnym na wyszukanie odpowiednich informacji, pojawia się kwestia złożoności obliczeniowej [1]. Potrzebne są zaawansowane narzędzia informatyczne i algorytmy, które poradzą sobie z wymienionymi trudnościami. Niezbędne jest ciągłe opracowywanie nowych lub usprawnionych rozwiązań, które będę w stanie poradzić sobie z coraz to większą ilością danych.

Przechowywanie danych w prostych strukturach, takich jak na przykład tabele, jest rozwiązaniem bardzo niekorzystnym ze względu na nakłady pamięci, jakie trzeba poświęcić na każdy rekord. Wspomniane tabele są strukturami na tyle nieefektywnymi, że wykorzystywanie ich do złożonych wyszukiwań, czy przechowywania informacji, gdzie liczba powielonych informacji jest duża i rośnie wraz ze wzrostem ilości rekordów, nie jest rozwiązaniem optymalnym.

Dlatego w dzisiejszych czasach dziedzina wiedzy, jaką jest eksploracja danych jest tematem aktualnym i wciąż rozważanym przez szereg naukowców. Coraz to nowe pomysły na transformacje prostych struktur do bardziej złożonych, pozwalających na szybkie wnioskowanie na temat różnego rodzaju relacji jest tematem, który wciąż można rozwijać i wyciągać cenne wnioski na temat już istniejącej bazy wiedzy w tym obszarze nauki.

1.1. Cel i zakres pracy

Praca ma na celu opracowanie oraz implementację asocjacyjnego modelu danych, do którego transformowane są klasyczne tabele pozwalając na szybkie (zwykle w czasie stałym O(1)) wnioskowanie na temat różnego rodzaju relacji automatycznie dostępnych w asocjacyjnych grafowych strukturach danych AGDS, bez konieczności żmudnego przeszukiwania danych w pętlach obliczeniowych. Zakres pracy obejmuje również porównanie szybkości działania operacji na tych strukturach z operacjami wykonywanymi na klasycznych strukturach tabelarycznych, wraz z analizą oraz wyprowadzeniem wniosków na temat złożoności obliczeniowej. Zbudowany system ma za zadanie również umożliwić zapis tych struktur w pamięciach trwałych. Działanie będzie zademonstrowane na wybranych zbiorach danych oraz zilustrowane graficznie.

Implementacja struktury grafowej oraz poszczególnych funkcji umożliwiających wyszukiwania odpowiednich relacji została zaimplementowana w języku C++ [2]. Środowisko IDE, wykorzystane przy implementacji kodu źródłowego, to QT Creator - bezpłatne wieloplatformowe środowisko programistyczne dla języków C++, JavaScript oraz QML, będące częścią SDK dla biblioteki Qt.

Dane wejściowe, które wykorzystano do analizy oraz przetworzenia za pomocą struktury grafowej, to zbiory danych dostępne w repozytorium danych UCI Machine Learning Repository [3]. Aktualnie zawiera 381 zbiorów danych, jako usługi dostępne dla społeczności zajmującej się tematem Machine Learning'u.

Programy Graphviz [4] oraz draw.io [5], w których wykonano wizualizacje dotyczące grafów to bezpłatne, dostępne online programy umożliwiające tworzenie zarówno prostych, jak i zaawansowanych schematów.

2. Asocjacyjna Struktura Danych AGDS

2.1. Graf AGDS – co to jest?

Grafowa asocjacyjna struktura danych AGDS (ang. Associative Grap Data Structure), to graf umożliwiający przechowywanie wartości danych i ich kombinacji, wraz z uproszczoną reprezentacją ich asocjacyjnego podobieństwa (ESIM), asocjacyjnego następstwa (ESEQ) i asocjacyjnego definiowania (EDEF), jakie występują pomiędzy nimi.

Struktura AGDS charakteryzuje się tym, że nie ma możliwości odwzorowania zależności czasowych, które są konieczne do zaprezentowania innego rodzaju asocjacyjnych powiązań. Dlatego też nie będą odwzorowane powiązania tłumiące nazywane ASUP, z kolei powiązania kontekstowe (ACON) stosowane w bardziej rozbudowanych strukturach neuronowych mogą być reprezentowane tylko w okrojonym stopniu. Kolejną istotną cechą struktury AGDS jest fakt, iż nie zawiera dynamicznych i zmiennych w czasie neuronów, wag i połączeń synaptycznych, przez co jest ona pasywna i statyczna. Węzły oraz krawędzie, odwzorowujące różne asocjacyjne zależności, reprezentują dane oraz kombinacje. Zazwyczaj krawędzie posiadają swoją wagę, której to wartość liczbowa jest określana jako siła połączenia dwóch węzłów. Często węzłom przypisuje się dodatkowe etykiety, mówiące o tym, co reprezentują w danej strukturze.

Struktury grafowe są wykorzystywane do przeszukiwania ich za pomocą odpowiednich algorytmów lub, gdy chcemy wykonywać asocjacyjne obliczenia, są przekształcane do postaci aktywnych grafów AANG należących do grupy asocjacyjnych systemów skojarzeniowych AAS. Struktury te wyróżniają się również tym, że modelują różnego rodzaju połączenia asocjacyjne, co w znaczny sposób upraszcza oraz przyśpiesza algorytmy, służące do wyszukiwania. Kombinacje oraz dane są uporządkowane dzięki odwzorowaniu tych relacji. Struktury grafowe są bardzo dobrą alternatywą dla innych metod przechowywania i reprezentacji informacji [1].

Struktury AGDS są niezwykle oszczędne ponieważ nie znajdują się w nich ani duplikaty informacji, ani duplikaty ich kombinacji. Informacje na temat podobieństw, różnic między nimi oraz korelacji są dostarczane właściwie automatycznie. Wadą tych struktur jest fakt, iż nie przechowują one niepowiązanych i niesparametryzowanych ciągów danych.

Od strony formalnej struktura grafowa AGDS to uporządkowana siódemka, gdzie VV, VR, VS oraz VC to zbiory wierzchołków, które reprezentują [1] :

- VV – pojedynczą wartość,
- VR - przedział wartości,

- VS - podzbiór wartości.

VC - kombinację wartości oraz zbiór krawędzi:

- ESIM – nieskierowanych, łączących asocjacyjnie podobne wierzchołki,
- ESEQ - skierowanych, łączących asocjacyjnie następne wierzchołki,
- EDEF - dwustronnie skierowanych łączących wierzchołek definiujący z definiowanym tak, że waga określa przejście pomiędzy wierzchołkami.

2.2. Złożoność obliczeniowa – tabele a grafy AGDS

W celu zdobycia jakichkolwiek informacji z danych reprezentowanych w tabeli, trzeba ją zwykle kilkukrotnie przeszukać w wielu pętlach, iterując po kolei po ogromnej ilości wierszy oraz po poszczególnych elementach w wierszach. Następnie na uzyskanych danych wykonywane są kolejne operacje, na przykład matematyczne, oraz sprawdzane zgodnie z zadaniem kolejne warunki. Tabele możemy sortować względem konkretnych parametrów, czy dodawać indeksy, które umożliwią nam posortowanie tabeli po kilku, a nawet po wszystkich parametrach.

Jeżeli chodzi o złożoność obliczeniową takich operacji, to operacje wyszukiwania będą kosztowały:

- ***O(n)* operacji na nieposortowanej tabeli**
- ***O(log n)* na posortowanej tabeli**
- ***O(n * log n)* posortowanie każdego parametru (n liczba wierszy)**

Tabele można sortować względem jednej kolumny lub przy sortowaniu względem dodatkowych kolumn, należy wprowadzić indeksację. Koszt samego wyszukania danych do dalszego przetworzenia jest zatem bardzo duży, tym większy, im większy jest zbiór danych.

Pomysł, aby połączyć ze sobą wszystkie istotne dane z odpowiednich atrybutów, wyeliminować redundancję oraz usunąć wszystkie duplikaty, wygląda na początek czegoś, co zaczyna przypominać użyteczną strukturę, jaką jest graf. Sama struktura jest jeszcze niejednoznaczna, więc trzeba dodać definiujące połączenia kontekstowe. Wszystkie dające się uporządkować elementy powinny być powiązane między sobą krawędziami, w zależności od podobieństwa, tak, aby dało się otrzymać od razu posortowaną strukturę obiektów danego rodzaju. Dzięki temu, można łatwo określić obiekty lub jakieś ich części o podobnych cechach.

Bez najmniejszego problemu i niskim kosztem obliczeniowym można określić wszystkie różnice i podobieństwa.

W momencie, w którym zaistnieje potrzeba uzyskania informacji na temat jakiegokolwiek rekordu, wartości konkretnego atrybutu lub innej kombinacji czy korelacji, informacje te będą praktycznie od razu dostępne, tzn. przy stałym koszcie ich wyszukania.

- **Oznacza to, że złożoność takiej operacji wynosi *O(1)***

Nie ma potrzeby iteracyjnego przeszukiwania danych, tak jak w przypadku tabel. Każdy rekord ma bezpośredni dostęp do wszystkich swoich wartości atrybutu, z kolei każda wartość atrybutu posiada bezpośrednie powiązanie i dostęp do obiektów, które definiuje. Bardzo ważne jest, aby dane powiązać ze sobą w taki sposób, by uzyskiwać do nich jak najszybszy dostęp. Asocjacyjne struktury danych AGDS, oprócz przechowywania danych, zawierają również informacje o relacjach, usuwają duplikaty oraz zapewniają szybkie wnioskowanie i eksploracje danych.

Analizując wszystkie możliwości oraz cechy grafowych struktur danych AGDS, można podsumować ich cechy i możliwości następująco:

- Grafy AGDS mogą być wykorzystywane jako forma reprezentacji oraz przechowywania danych.
- Służą do znajdowania i przechowywania asocjacyjnych relacji pomiędzy danymi, obiektami i zdefiniowanymi na ich podstawie informacjami.
- Są wykorzystywane do łatwego oraz szybkiego znajdywania powiązanych danych oraz ich kombinacji, korelacji, podobieństw i różnic.
- Poprzez formę w jakiej przetrzymywane są dane (brak duplikatów), bez strat kompresują i kontekstowo wiążą informacje, szczególnie w przypadku dużych zbiorów danych.

2.3. Konwersja danych do postaci struktury grafu AGDS

Dla małych zbiorów danych, porównując tabele oraz grafową strukturę danych AGDS można odnieść wrażenie, że przechowywanie informacji w grafie jest niekorzystne, gdyż wymaga transformacji i ze względu na ograniczoną ilość duplikatów w takich zbiorach może być bardziej kosztownym sposobem przechowywania danych. Oprócz podstawowych wartości, czyli węzłów grafu, struktury grafowe zawierają również powiązania pomiędzy tymi wartościami, które nazywamy krawędziami. Jeżeli głównym celem zastosowania grafu jest tylko przechowywanie danych, w takim przypadku jest to rozwiązanie rzeczywiście niekorzystne. Co do samej wielkości zbioru danych, im mniejszy zbiór, tym bardziej

nieoptymalnym jest przechowywanie go w tak złożonej strukturze, jaką jest graf. Pełną użyteczność grafowych zbiorów danych można dostrzec w sytuacji, kiedy potrzebne jest wyodrębnienie ważnych relacji pomiędzy informacjami.

Prawdą jest, że transformacja danych do struktury grafowej jest kosztowna (pesymistyczna złożoność takiej operacji trwa O(n log n), jednak dzięki temu uzyskujemy bogatą funkcjonalność, której oszczędność ujawnia się szczególnie przy reprezentacji ogromnych zbiorów danych. Dodatkowy zysk osiągamy dzięki braku duplikacji danych oraz ich kombinacji, a także agregację powiązań.

Konwersję informacji przechowywanych pierwotnie w klasycznych strukturach tabelarycznych do struktury grafowej AGDS zaczynamy od przekształcenia nazw atrybutów, nazw rekordów oraz wszystkich wartości w węzły grafu AGDS. Węzły reprezentujące wartości oraz rekordy są połączone krawędziami EDEF. Duplikaty wartości atrybutów oraz rekordów są automatycznie usuwane, dzięki temu wyszukiwanie podobieństw oraz różnic jest bardzo proste. W przypadku użycia kilku powiązanych ze sobą tabel, powiązania te są oznaczane przy użyciu krawędzi EDEF. Wartości danych znajdujących się w konkretnych kolumnach są połączone również wyżej wymienionymi krawędziami, z węzłami reprezentującymi konkretny atrybut. Węzły wartości atrybutów reprezentujące sąsiednie wartości są połączone krawędziami ESIM. Jeżeli jest to tylko możliwe, wartości te są automatycznie sortowane, przez co otrzymuje się uporządkowane i niezduplikowane listy wartości dla konkretnych atrybutów. W przypadku, gdy kolejność rekordów ma jakieś znaczenie, są one połączone krawędziami ESEQ. Struktura grafowa AGDS odwzorowuje wszystkie znane podstawowe relacje pomiędzy rekordami oraz jego danymi.

Węzły, które są wspólnymi wartościami atrybutów, łączą się z rożnymi węzłami reprezentującymi rekordy. Dzięki temu od razu można wskazać wzorce, które zawierają mniejsze lub większe wartości dla konkretnych atrybutów. Jeżeli w zbiorze danych znajdują się klasy, które są również cechą rekordów, wtedy są traktowane w grafie tak samo jak atrybuty. Dzięki temu, iż w grafie zawarte są posortowanie listy konkretnych atrybutów, ułatwione jest znajdywanie kolejnych i poprzednich wartości oraz podobnych rekordów, nie wykonując przy tym dodatkowych operacji takich jak sortowanie.

Konwersja zbiorów danych z tabel do postaci grafowych struktur AGDS [6,7]:

- sortuje wszystkie wartości atrybutów oraz rekordy,
- usuwa duplikaty,
- umożliwia szybkie określenie podobieństw, korelacji, kolejności, różnic pomiędzy rekordami w zależności od podanych parametrów lub innych kryteriów zdefiniowanych przez użytkownika,
- umożliwia łatwe wykrywanie i eksplorację danych.

3. Eksperymenty

3.1. Zaimplementowany graf AGDS

Pierwszym krokiem podczas tworzenia programu umożliwiającego wnioskowanie na temat relacji jest zaimplementowanie asocjacyjnego modelu danych.

Na samym początku został stworzony obiekt typu parametr, który będzie stanowił korzeń dla zaimplementowanego grafu. Na cały graf zostanie utworzona tylko jedna instancja tego obiektu oraz nie będzie ona identyfikowana z żadną konkretną wartością. Co najważniejsze zawierać będzie ona listę wszystkich atrybutów, wczytaną według kolejności podanej w danych wejściowych.

Danymi wejściowymi dla zbudowanego grafu są dane tabelaryczne. Dane te są wczytywane linijka po linijce, tworząc asocjacyjną strukturę danych. W pierwszym wierszu znajdują się nazwy wszystkich atrybutów. Każde kolejne wyrażenie jest wczytywane i automatycznie tworzone są kolejne atrybuty, które będą identyfikowane przez nazwę lub poprzez kolejność.

Przy wczytywaniu każdego nowego wiersza tworzony jest kolejny obiekt rekordu. Każdy nowy rekord jest identyfikowany poprzez swój numer. Ze względu na fakt, iż wiersze są wczytywane po kolei, kolejność jest dobrym identyfikatorem obiektów typu rekord.

Każdy wiersz składa się z kolejnych elementów dostępnych w kolumnach. Każdy taki element wiersza jest identyfikowany jako nowy obiekt typu wartość. Bardzo ważne jest, że jeżeli istnieje już obiekt typu wartość, identyfikowany przez identyczną wartość, nowy obiekt nie jest tworzony. Oczywiste jest również, że najważniejszym elementem tego obiektu, poprzez który będzie on rozpoznawany w przyszłości, jest właśnie wczytana wartość. Nowo stworzony obiekt trafia do listy wartości w obiekcie rekordu, do którego należy. Wszystkie obiekty typu wartość z jednego wiersza trafiają na listę znajdującą się w jednym rekordzie, do którego przynależą. Oprócz tego kolejne obiekty typu wartość w jednym wierszu są identyfikowane jako wartości reprezentujące konkretne atrybuty. Stąd przy tworzeniu kolejnych obiektów wartości trafiają na konkretne listy wartości znajdujące się we wcześniej stworzonych obiektach atrybutów. Kolejność w wierszu ma znaczenie, ponieważ jak już wcześniej zostało wspomniane, w tej samej kolejności, w jakiej zostały wczytywane obiekty atrybutów, w takiej samej kolejności obiekty wartości są identyfikowane z nimi.

Po wczytaniu wszystkich wierszy, następuje sortowanie obiektów typu wartość w listach znajdujących się w obiektach atrybutów. Oprócz tego, w obiekcie każdego atrybutu zostają ustawione wskaźniki wskazujące na obiekty typu wartość, reprezentujące najmniejszą oraz największą wartość liczbową w atrybucie. Obiekt typu atrybut posiada jeszcze element reprezentujący różnicę pomiędzy najmniejszą, a największą wartością znajdującą się w danym

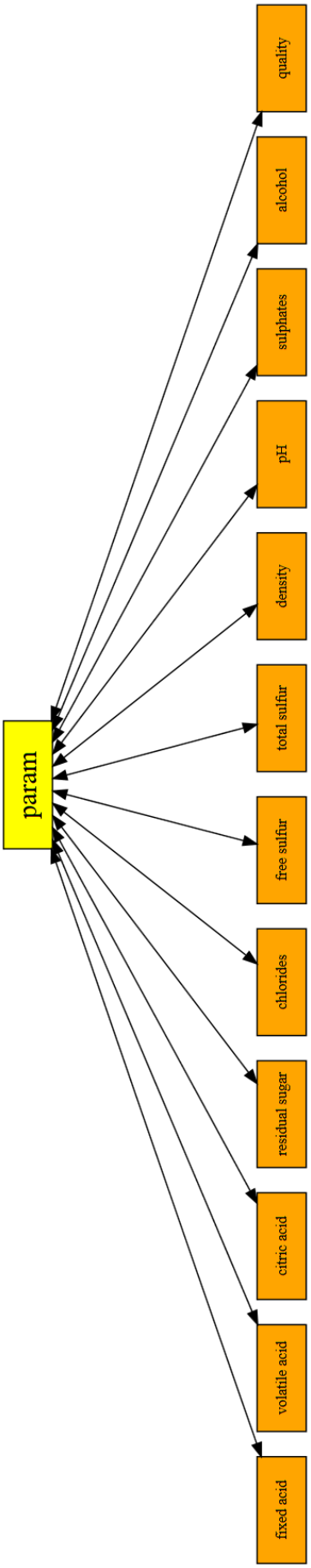
atrybucie, wykorzystywaną do wyznaczenia wagi połączeń pomiędzy obiektami reprezentującymi sąsiednie wartości.

Podsumowując, zaimplementowany graf składa się z następujących elementów:

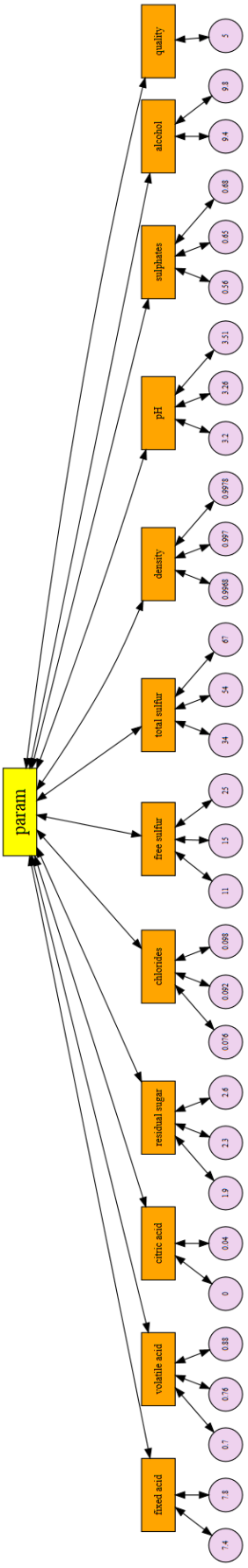
- Jednej instancji obiektu typu parametr, stanowiącej korzeń grafu oraz zawierającej listę wszystkich obiektów atrybutów.
- Obiektów typu atrybut zawierających nazwę atrybutu, posortowaną listę obiektów typu wartość należących do konkretnego atrybutu, wskaźnik do obiektu typu parametr, obiekty reprezentujące najmniejszą oraz największą wartość w danym atrybucie, zakres wartości.
- Niezduplikowanych obiektów typu wartość zawierających element reprezentujący wartość liczbową, wskaźnik do obiektu typu atrybut, do którego należy konkretna wartość, listę obiektów typu rekord, wskaźniki do poprzedzającego oraz następnego obiektu typu wartość, reprezentujące obiekty o mniejszej oraz większej wartości liczbowej. Pomocniczo w obiektach typu wartość znajdują się również dwa inne elementy nazwane „waga” oraz „oznaczony”. Będą one wykorzystywane przy liczeniu relacji alternatywy, koniunkcji oraz podobieństwa.
- Obiektów typu rekord zawierających numer rekordu, poprzez który będą identyfikowane, oraz listy obiektów typu wartość, które reprezentują konkretny rekord. Pomocniczo w obiekcie typu rekord znajdują się takie elementy jak „podobieństwo” oraz „powtórzenia”.

Tabela 1. Fragment zbioru danych wejściowych dla zbioru RedWine.

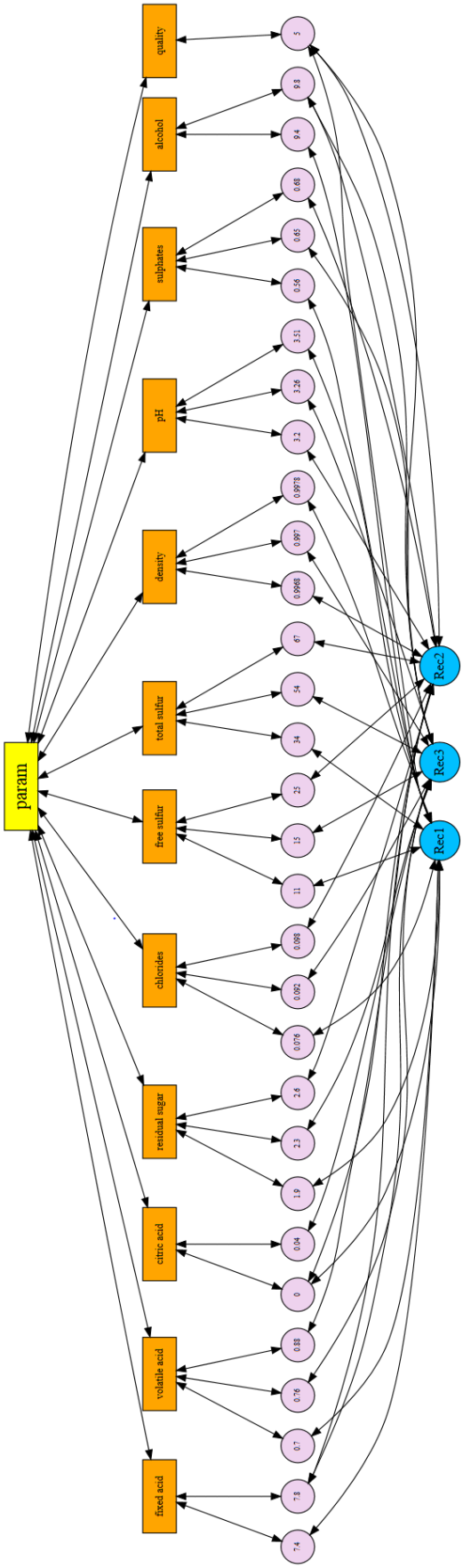
fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	q
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5
7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5
7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7
7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7
7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
6.7	0.58	0.08	1.8	0.097	15	65	0.9959	3.28	0.54	9.2	5
7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
5.6	0.615	0	1.6	0.089	16	59	0.9943	3.58	0.52	9.9	5
7.8	0.61	0.29	1.6	0.114	9	29	0.9974	3.26	1.56	9.1	5
8.9	0.62	0.18	3.8	0.176	52	145	0.9986	3.16	0.88	9.2	5
8.9	0.62	0.19	3.9	0.17	51	148	0.9986	3.17	0.93	9.2	5
8.5	0.28	0.56	1.8	0.092	35	103	0.9969	3.3	0.75	10.5	7
8.1	0.56	0.28	1.7	0.368	16	56	0.9968	3.11	1.28	9.3	5
7.4	0.59	0.08	4.4	0.086	6	29	0.9974	3.38	0.5	9	4
7.9	0.32	0.51	1.8	0.341	17	56	0.9969	3.04	1.08	9.2	6
8.9	0.22	0.48	1.8	0.077	29	60	0.9968	3.39	0.53	9.4	6
7.6	0.39	0.31	2.3	0.082	23	71	0.9982	3.52	0.65	9.7	5
7.9	0.43	0.21	1.6	0.106	10	37	0.9966	3.17	0.91	9.5	5
8.5	0.49	0.11	2.3	0.084	9	67	0.9968	3.17	0.53	9.4	5
6.9	0.4	0.14	2.4	0.085	21	40	0.9968	3.43	0.63	9.7	6
6.3	0.39	0.16	1.4	0.08	11	23	0.9955	3.34	0.56	9.3	5
7.6	0.41	0.24	1.8	0.08	4	11	0.9962	3.28	0.59	9.5	5
7.9	0.43	0.21	1.6	0.106	10	37	0.9966	3.17	0.91	9.5	5
7.1	0.71	0	1.9	0.08	14	35	0.9972	3.47	0.55	9.4	5
7.8	0.645	0	2	0.082	8	16	0.9964	3.38	0.59	9.8	6
6.7	0.675	0.07	2.4	0.089	17	82	0.9958	3.35	0.54	10.1	5
6.9	0.685	0	2.5	0.105	22	37	0.9966	3.46	0.57	10.6	6
8.3	0.655	0.12	2.3	0.083	15	113	0.9966	3.17	0.66	9.8	5
6.9	0.605	0.12	10.7	0.073	40	83	0.9993	3.45	0.52	9.4	6
5.2	0.32	0.25	1.8	0.103	13	50	0.9957	3.38	0.55	9.2	5
7.8	0.645	0	5.5	0.086	5	18	0.9986	3.4	0.55	9.6	6
7.8	0.6	0.14	2.4	0.086	3	15	0.9975	3.42	0.6	10.8	6
8.1	0.38	0.28	2.1	0.066	13	30	0.9968	3.23	0.73	9.7	7
5.7	1.13	0.09	1.5	0.172	7	19	0.994	3.5	0.48	9.8	4
7.3	0.45	0.36	5.9	0.074	12	87	0.9978	3.33	0.83	10.5	5
7.3	0.45	0.36	5.9	0.074	12	87	0.9978	3.33	0.83	10.5	5
8.8	0.61	0.3	2.8	0.088	17	46	0.9976	3.26	0.51	9.3	4



Rysunek 1. Fragment asocjacyjnego grafu przedstawiającego połączenie obiektów atrybutów z obiektem reprezentującym korzeń dla zbioru RedWine.



Rysunek 2. Fragment asocjacyjnego grafu przedstawiającego połączenie obiektów wartości połączonymi z atrybutami oraz obiektów atrybutów z obiektem reprezentującym korzeń dla zbioru RedWine.



Rysunek 3. Fragment asocjacyjnego grafu przedstawiającego połączenie pierwszych trzech rekordów z wartościami, obiektów wartości z atrybutami oraz obiektów atrybutów z obiektem reprezentującym korzeń dla zbioru RedWine.

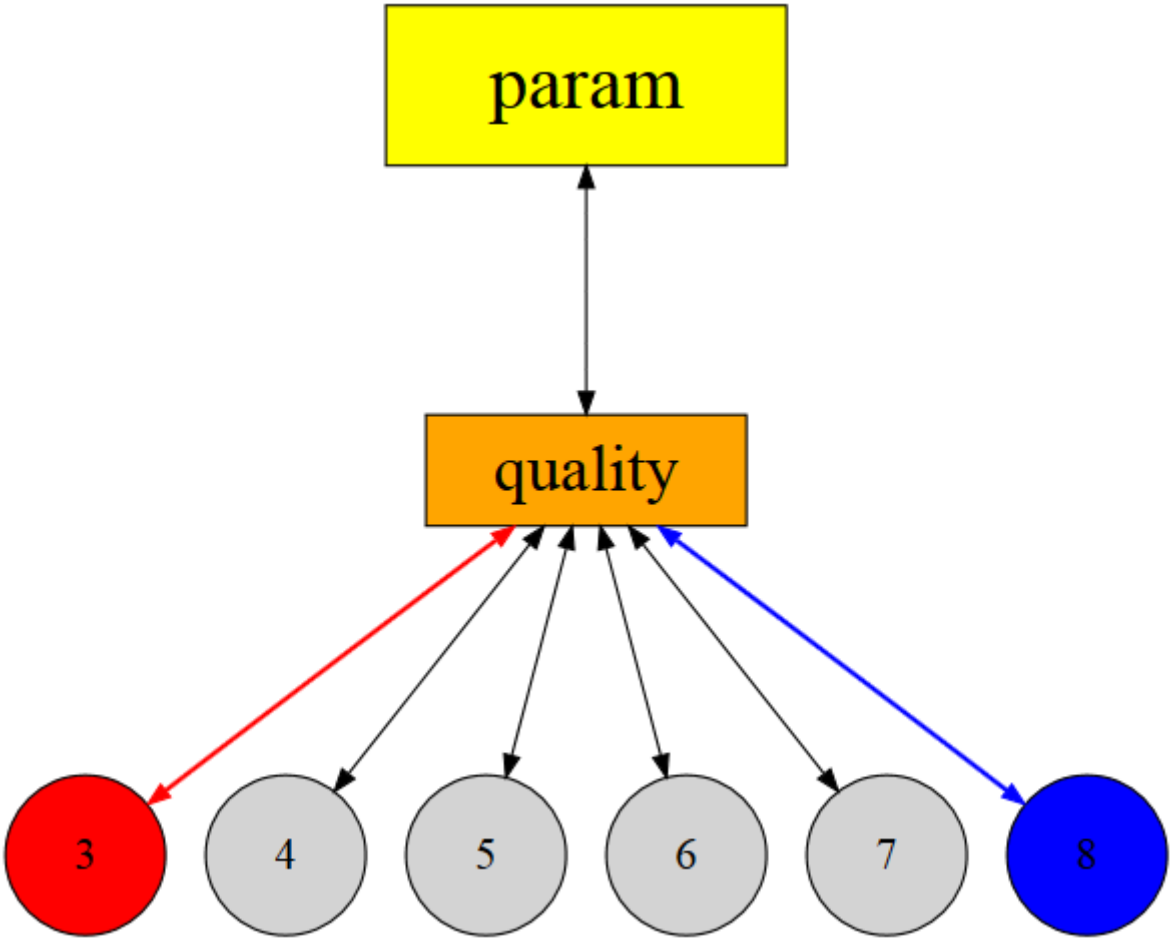
3.2. Minimum, maksimum i zakres

3.2.2. Minimum, maksimum i zakres w grafie

Konstrukcja zbudowanego grafu pozwala na bardzo szybkie, właściwie natychmiastowe, odnajdywanie prostych informacji o zbiorze danych. W klasie atrybutu znajduje się posortowana lista obiektów wartości należących do konkretnego atrybutu. Oprócz tego, elementami tej klasy są takie informacje, jak minimalna wartość ze zbioru, maksymalna wartość oraz zakres wartości.

Jeżeli chodzi o sposób znalezienia powyższych informacji, to minimum oraz maksimum uzyskujemy wprost pobierając pierwszy oraz ostatni element posortowanej listy obiektów wartości znajdującej się w obiekcie atrybutu. Co do zakresu, jest on po prostu wyliczany jako różnica wartości reprezentujących minimum oraz maksimum ze zbioru.

Złożoność obliczeniowa takiego algorytmu wynosi O(1), co oznacza, że niezależnie od wielkości zbioru wykonuje się stałą liczbę operacji dominujących.



Rysunek 4. Fragment asocjacyjnego grafu dla zbioru RedWine przedstawiający atrybut „quality” oraz jego posortowany zbiór obiektów wartości z zaznaczonymi wartościami minimum oraz maksimum.

3.2.3. Minimum, maksimum i zakres w tabeli

Jeżeli chodzi o wyszukiwanie takich informacji, jak minimum czy maksimum atrybutu w danych umieszczonych w tabeli, operacja nie jest tak prosta, jak w przypadku grafu AGDS. Na przykładzie zbioru *RedWine* zostanie pokazane, jak bardzo oszczędnym jest uzyskanie tych informacji z grafu.

W przykładzie powyżej, pokazano jako wygląda lista posortowanych obiektów wartości dla atrybutu „quality". Można zauważyć, że kontener ten zawiera sześć różnych wartości liczbowych posortowanych w kolejności rosnącej. Wydawać by się mogło, że dla takiego zbioru danych wyszukanie minimum i maksimum w tabeli nie będzie bardzo kosztowne.

Jeżeli chodzi o tabelę zawiera ona aż 1599 pól reprezentujących atrybut „quality". Są one nieposortowane, a ilość duplikatów jest tak ogromna, że konwersja z 1599 do 6 posortowanych pól jest dużą oszczędnością.

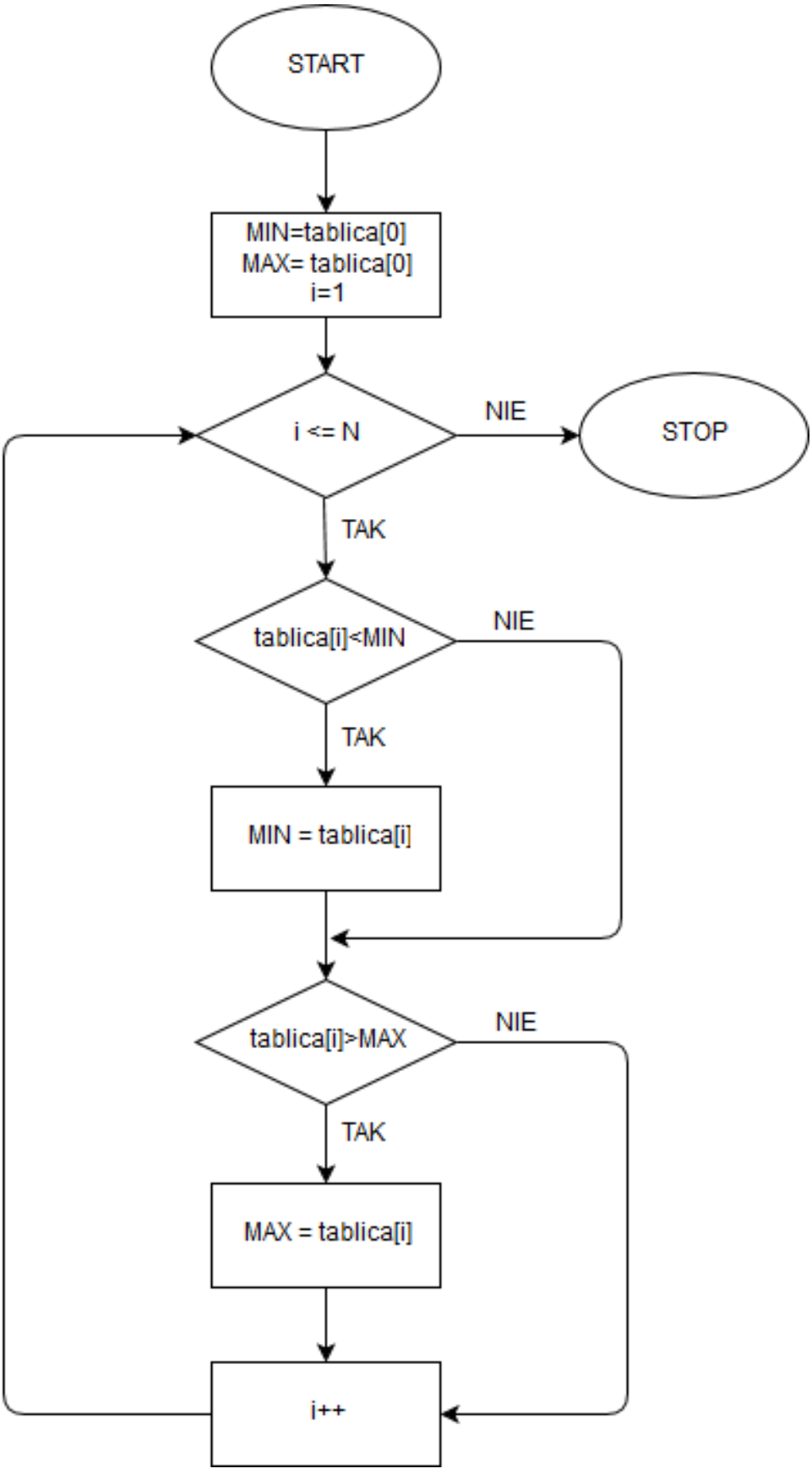
Aby znaleźć minimum i maksimum w zbiorze wartości znajdujących się w tablicy, należy wykonać iteracje po wszystkich elementach tablicy. W przypadku zbioru *RedWine,* należy wykonać w pętli 1599 iteracji, w przypadku większych zbiorów – więcej.

W każdej iteracji należy sprawdzić dwa warunki:

- Czy obecna wartość minimum jest większa niż i-ty element tablicy
- Czy obecna wartość maksimum jest mniejsza niż i-ty element tablicy

Jeżeli któryś z warunków jest prawdziwy, wtedy odpowiednio ustawiana jest nowa wartość minimum, bądź maksimum.

Złożoność obliczeniowa takiego algorytmu wynosi O(n), co oznacza, że w zależności od wielkości zbioru należy wykonać przynajmniej n operacji dominujących. Natomiast operacja dominująca, to taka operacja *o*, że liczba wszystkich operacji wykonanych przez algorytm jest nie większa niż $c \times$ *liczba operacji o* dla pewnej stałej $c \in N$. W szczególności oznacza to, że sama operacja dominująca musi składać się ze stałej (niezależnej od rozmiaru danych) liczby podoperacji.



Rysunek 5. Algorytm wybierania minimum i maksimum z tablicy.

Tabela 2. Czas wyszukiwania minimum, maksimum oraz zakresu w trzech różnych zbiorach.

	CZAS W MIKROSEKUNDACH					
	ForestFires		RedWine		OnlineNewsPopularity	
Nr atrybutu	Graf	Tabela	Graf	Tabela	Graf	Tabela
1	6	12	6	33	4	1625
2	2	11	2	27	1	1561
3	1	10	1	27	1	1508
4	2	10	1	27	1	1524
5	1	10	2	27	1	1028
6	2	11	1	27	1	718
7	2	12	1	27	0	714
8	2	10	1	28	1	710
9	1	11	2	27	1	724
10	1	15	1	27	0	742
11	1	18	2	28	1	755
12	1	15	2	26	1	833
13	1	18			1	704
14					1	738
15					1	820
16					1	783
17					1	1134
18					1	1064
19					0	694
20					1	615
21					1	610
22					0	618
23					0	609
24					1	603
25					1	672
26					1	693
27					1	615
28					1	618
29					1	604
30					1	633
31					1	859
32					0	1296
33					1	1446
34					1	1633
35					1	1945
36					1	1763
37					0	624
38					1	607
39					11	609
40					2	615

41					1	600
42					1	638
43					1	611
44					1	601
45					2	603
46					2	605
47					2	602
48					2	643
49					2	608
50					2	604
51					1	610
52					1	637
53					1	630
54					1	611
55					1	630
56					1	602
57					2	608
58					2	630
59					2	706
60					2	607

W powyższej tabeli przedstawiono wyniki czasowe dotyczące trwania wyszukiwania w asocjacyjnym grafie oraz w tabelach informacji takich jak minimum, maksimum oraz zakres dla każdego z atrybutów reprezentujących konkretne zbiory. Zbiory danych, które zostały wykorzystane to:

- *ForestFires* zawierający 13 atrybutów oraz 517 rekordów,
- *RedWine* zawierający 12 atrybutów oraz 1599 rekordów,
- *OnlineNewsPopularity* zawierający 60 atrybutów oraz 39 644 rekordy.

Należy zwrócić uwagę na fakt, że czas wykonywania operacji we wszystkich trzech zbiorach jest taki sam, jeżeli chodzi o operacje wykonywane na grafowej strukturze danych. Operacje te dzieją się tak naprawdę natychmiastowo.

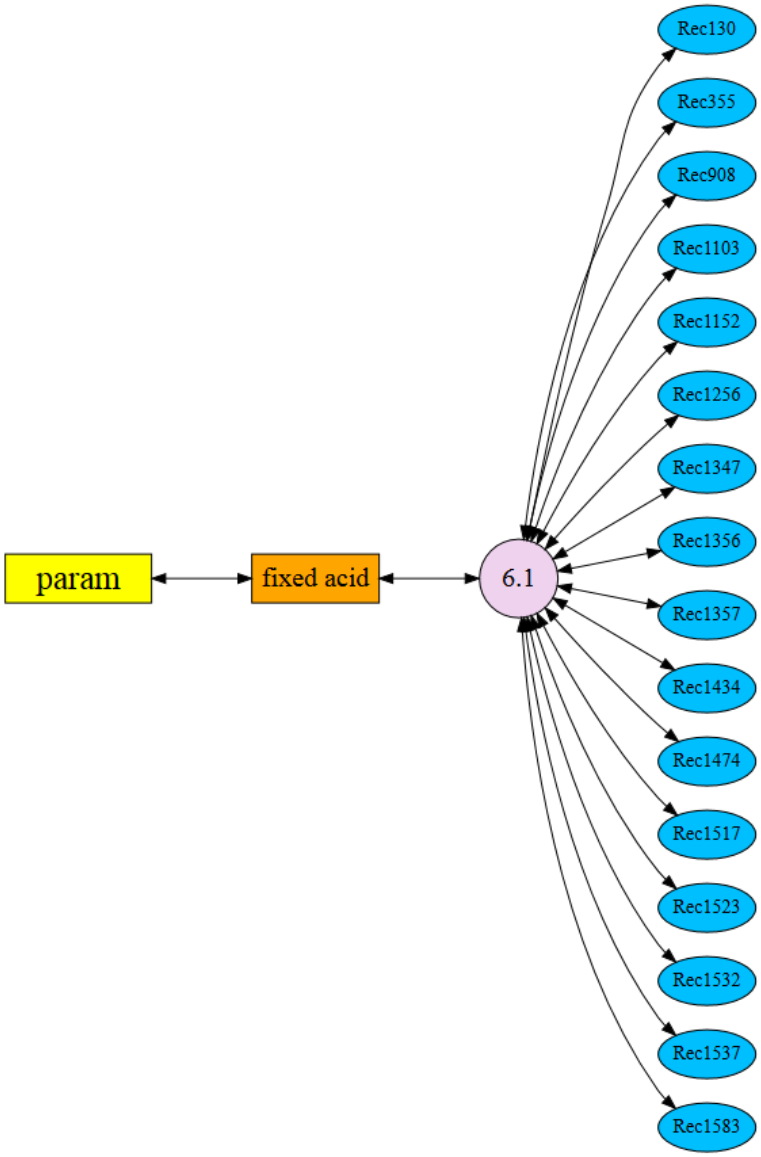
Jeżeli rozpatrzymy operacje wykonywane na tabelach widzimy wyraźnie, że oprócz tego, iż same operacje trwają dłużej niż w przypadku grafów, to im większy jest zbiór danych, tym czas tych operacji są coraz większy.

3.3. Wyszukiwanie rekordów o określonej wartości

3.3.1. Wyszukiwanie rekordów o określonej wartości w grafie

Jeżeli chodzi o wyszukanie takich informacji jak listy rekordów, które mają konkretną wartość we wskazanym atrybucie, to w grafie jest to równie proste jak poprzednia operacja. Dzięki temu, że obiekt reprezentujący wartość posiada kontener zawierający listę wszystkich rekordów powiązanych z nim, to informację o rekordach reprezentowanych przez określoną wartość w atrybucie, otrzymujemy od ręki.

Złożoność obliczeniowa takiego algorytmu wynosi O(1), co oznacza, że niezależnie od wielkości zbioru należy wykonać stałą liczbę operacji dominujących.

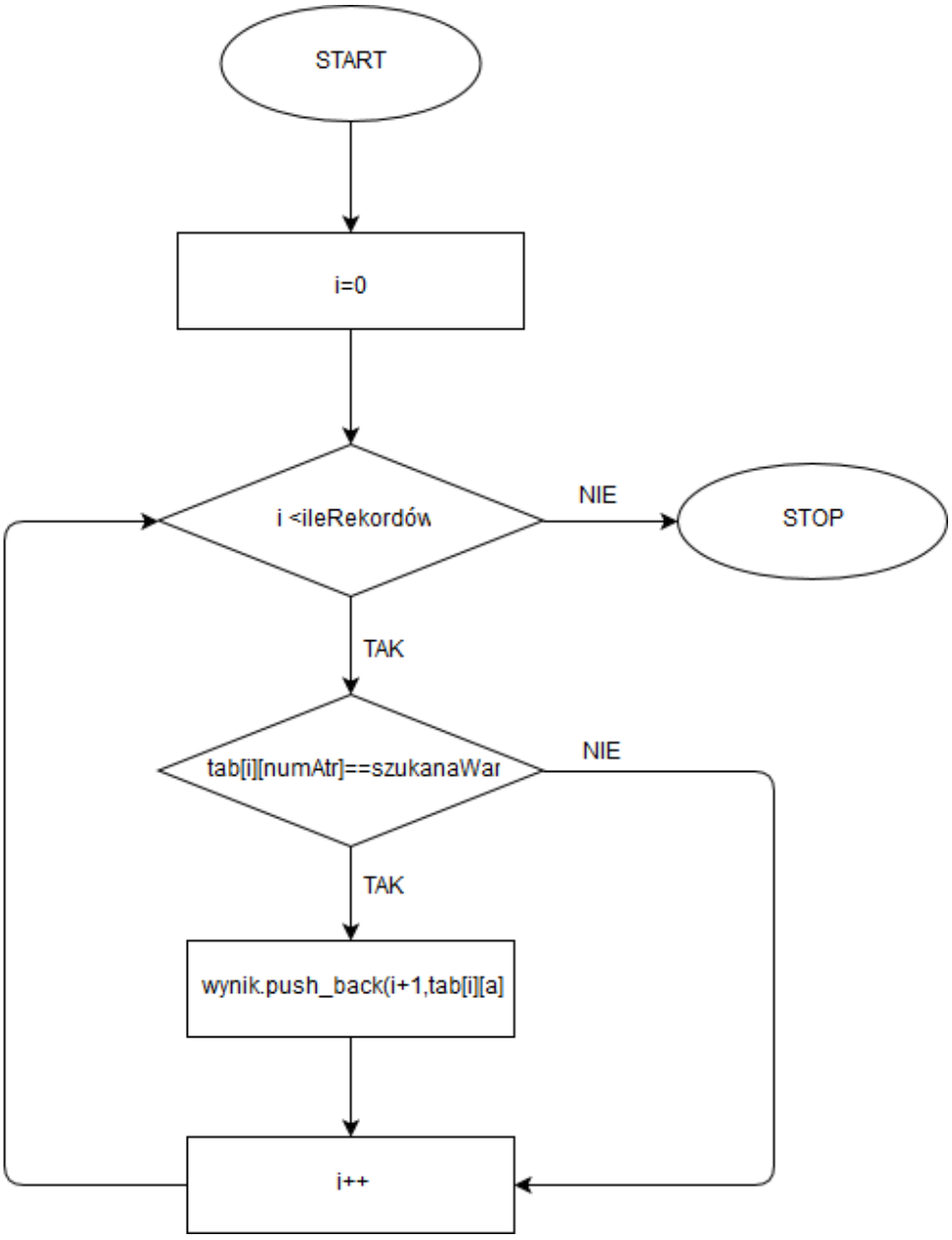


Rysunek 6. Fragment asocjacyjnego grafu dla zbioru RedWine przedstawiający dla wartości 6.1 i atrybutu „fixed acid" wszystkie powiązane rekordy.

3.3.2. Wyszukiwanie rekordów o określonej wartości w tabeli

Aby wyszukać informacje o wszystkich rekordach posiadających określoną wartość we wskazanym atrybucie, tak naprawdę jest koniecznym przeiterowanie po całej tabeli. Z każdą iteracją należy sprawdzić, czy szukana wartość znajduje się we wskazanym wierszu. Jeżeli tak, numer rekordu oraz cały wiersz zostaje zapisany do końcowej listy wyników.

Złożoność obliczeniowa takiego algorytmu wynosi O(n), co oznacza, że w zależności od wielkości zbioru należy wykonać przynajmniej n operacji dominujących. Operacjami dominującymi są tutaj sprawdzenie, czy wartość znajduje się we wskazanym wierszu, a jeżeli tak, to wpisanie rekordu do listy wyników.



Rysunek 7. Algorytm wyszukiwania rekordów o określonej wartości w tabeli.

3.3.3. Wyniki czasowe dla wyszukiwania rekordów o określonej wartości

Tabela 3. Czas wyszukiwania rekordów o określonej wartości we wszystkich atrybutach ze zbioru ForestFires.

Nazwa atrybutu	ForestFires
X	34, 31, 23, 37, 13, 35, 25, 25, 6
	101, 128, 96, 150, 62, 209, 101, 103, 41
Y	18, 25, 76, 46, 26, 2, 3
	83, 115, 345, 207, 129, 8, 30
month	2, 9, 21, 5, 2, 8, 13, 70, 64, 7, 2, 4
	23, 48, 100, 35, 24, 43, 66, 294, 277, 39, 8, 30
day	31, 25, 22, 24, 33, 31, 36
	133, 113, 96, 120, 147, 142, 162
FFMC	2, 1, 1, 2, 2, 2, 1, 2, 2, 3, 2, 1, 1, 2, 1, 1, 1, 3, 1, 2, 2, 1, 1, 2, 3, 1, 1, 2, 2, 1, 1, 1, 1, 1, 2, 1, 3, 2, 1, 3, 1, 1, 2, 2, 2, 2, 3, 1, 2, 3, 2, 5, 5, 3, 3, 3, 4, 3, 1, 4, 8, 5, 5, 1, 5, 4, 10, 8, 4, 3, 2, 10, 4, 3, 6, 7, 3, 1, 3, 5, 2, 3, 2, 2, 3, 5, 2, 7, 2, 1, 2, 4, 1, 1, 2, 6, 2, 1, 2, 4, 1, 1, 1, 2, 2, 2
	9, 8, 8, 22, 8, 8, 24, 25, 24, 27, 8, 8, 8, 21, 20, 20, 7, 23, 21, 22, 24, 7, 7, 8, 25, 7, 7, 20, 7, 7, 21, 7, 7, 7, 23, 7, 25, 24, 7, 25, 21, 7, 24, 22, 8, 21, 24, 21, 23, 24, 20, 35, 36, 30, 27, 27, 30, 29, 24, 35, 50, 32, 35, 7, 34, 31, 59, 48, 32, 26, 22, 57, 31, 26, 41, 38, 24, 8, 26, 34, 25, 27, 20, 25, 26, 34, 26, 41, 22, 7, 24, 31, 8, 21, 20, 38, 21, 8, 25, 27, 21, 8, 21, 20, 26, 21
DMC	2, 1, 2, 1, 2, 2, 2, 2, 1, 1, 1, 1, 1, 2, 2, 2, 2, 1, 1, 2, 2, 1, 1, 2, 1, 1, 1, 2, 2, 2, 2, 1, 1, 2, 2, 3, 3, 2, 1, 1, 2, 1, 1, 1, 1, 2, 1, 1, 2, 2, 4, 1, 2, 2, 2, 2, 2, 1, 2, 2, 1, 3, 2, 2, 1, 1, 3, 3, 1, 2, 1, 2, 1, 1, 1, 1, 1, 2, 2, 1, 2, 2, 2, 1, 2, 2, 3, 2, 2, 2, 2, 1, 3, 2, 2, 3, 1, 1, 2, 2, 18, 1, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 2, 4, 2, 2, 2, 2, 1, 4, 1, 1, 2, 2, 2, 2, 3, 3, 2, 1, 3, 2, 1, 2, 2, 2, 2, 3, 3, 3, 1, 2, 2, 1, 2, 0, 1, 0, 2, 1, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 2, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1
	2, 2, 6, 2, 2, 9, 6, 2, 2, 2, 3, 10, 2, 2, 2, 2, 6, 2, 2, 6, 6, 2, 2, 2, 2, 5, 2, 2, 5, 5, 2, 2, 2, 6, 2, 8, 7, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 5, 6, 6, 7, 2, 6, 6, 2, 2, 5, 2, 2, 6, 2, 7, 6, 6, 6, 2, 7, 6, 2, 2, 2, 6, 2, 2, 1, 2, 2, 5, 2, 2, 5, 5, 2, 2, 6, 6, 7, 5, 5, 5, 2, 2, 7, 6, 2, 7, 2, 2, 2, 6, 5, 2, 6, 2, 5, 2, 6, 6, 6, 6, 2, 6, 6, 2, 9, 6, 5, 2, 6, 2, 7, 5, 2, 2, 7, 7, 5, 7, 7, 2, 2, 7, 7, 6, 6, 6, 5, 5, 7, 7, 7, 2, 2, 6, 2, 7, 6, 7, 2, 8, 7, 6, 7, 6, 7, 2, 2, 2, 5, 2, 7, 2, 7, 2, 6, 6, 8, 7, 2, 6, 2, 8, 3, 2, 7, 7, 6, 2, 2, 2, 2, 6, 2, 2, 5, 6, 6, 2, 5, 8, 7, 2, 5, 5, 2, 6, 7, 2, 5, 6, 6, 8, 7, 2, 6, 2, 2, 2, 2, 5, 6, 2, 2, 6, 10
DC	1, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 8, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1
	3, 2, 2, 2, 3, 6, 2, 14, 2, 6, 2, 5, 2, 2, 2, 2, 2, 5, 6, 2, 2, 2, 2, 6, 6, 2, 6, 2, 5, 6, 2, 2, 2, 6, 7, 2, 2, 6, 2, 2, 2, 2, 2, 5, 5, 2, 7, 7, 2, 2, 2, 2, 2, 6, 2, 2, 6, 2, 6, 6, 2, 2, 2, 7, 2, 2, 2, 5, 1, 2, 2, 2, 2, 2, 2, 2, 6, 2, 5, 2, 2, 2, 2, 1, 2, 2, 2, 2, 6, 7, 6, 6, 2, 2, 5, 2, 6, 2, 2, 2, 6, 2, 2, 2, 1, 5, 2, 7, 6, 6, 6, 2, 15, 3, 8, 2, 15, 2, 7, 2, 7, 14, 2, 2, 2, 6, 2, 7, 6, 14, 5, 6, 7, 2, 7, 15, 6, 6, 6, 6, 6, 7, 9, 8, 6, 7, 7, 7, 6, 7, 7, 2, 2, 3, 6, 8, 7, 6, 6, 6, 8, 9, 2, 6, 8, 7, 6, 6, 7, 2, 6, 3, 6, 2, 7, 6, 8, 2, 2, 6, 6, 6, 2, 2, 7, 6, 3, 2, 2, 7, 7, 8, 8, 12, 8, 10, 9, 11, 10, 11, 3, 9, 8, 11, 9, 6, 2, 3, 7, 3, 6, 2, 6, 2, 7, 2, 2, 6, 2
ISI	1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 2, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 3, 3, 1, 1, 1, 1, 1, 3, 3, 1, 0, 0, 2, 1, 1, 2, 0, 1, 1, 1, 1, 2, 2, 1, 1, 1, 2, 2, 2, 0, 1, 3, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 2, 1, 1, 2, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0
	3, 7, 2, 7, 2, 2, 2, 7, 2, 6, 2, 2, 7, 6, 6, 6, 6, 7, 2, 2, 7, 6, 7, 6, 3, 7, 3, 6, 7, 5, 6, 6, 2, 6, 7, 8, 7, 12, 13, 6, 7, 2, 7, 8, 13, 14, 8, 2, 7, 10, 7, 7, 10, 2, 6, 10, 7, 2, 13, 9, 8, 6, 7, 8, 10, 10, 2, 7, 15, 2, 2, 8, 8, 8, 8, 9, 6, 2, 8, 6, 6, 6, 3, 6, 6, 2, 7, 7, 8, 2, 2, 2, 6, 7, 6, 7, 8, 6, 6, 10, 2, 2, 6, 6, 7, 2, 3, 2, 6, 8, 8, 8, 8, 7, 8, 2, 2, 3, 2
temp	0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 2, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 2, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 2, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1

	3, 2, 9, 2, 8, 3, 6, 2, 6, 2, 6, 2, 6, 6, 7, 6, 2, 2, 2, 2, 6, 2, 6, 2, 2, 7, 2, 6, 7, 6, 2, 6, 6, 2, 6, 6, 3, 6, 2, 6, 2, 6, 2, 2, 5, 6, 7, 7, 7, 2, 2, 7, 6, 6, 2, 6, 7, 2, 2, 6, 7, 7, 6, 6, 6, 6, 7, 2, 6, 6, 2, 6, 8, 7, 7, 2, 7, 6, 6, 6, 8, 7, 7, 8, 7, 7, 2, 8, 7, 7, 6, 7, 7, 6, 7, 6, 8, 6, 7, 7, 7, 9, 6, 7, 6, 8, 7, 7, 8, 6, 8, 8, 8, 6, 6, 7, 7, 7, 6, 6, 8, 6, 6, 8, 7, 7, 6, 6, 2, 6, 7, 8, 8, 7, 2, 6, 7, 9, 7, 2, 7, 7, 7, 2, 7, 7, 7, 2, 7, 6, 6, 2, 3, 6, 6, 2, 2, 2, 7, 2, 6, 3, 7, 2, 6, 6, 2, 2, 3, 6, 2, 2, 6, 6, 5, 2, 6, 2, 6, 2, 3, 2, 3, 6, 2, 7, 2, 3, 6, 2, 2, 2
RH	1, 0, 0, 2, 0, 1, 1, 2, 2, 1, 4, 2, 2, 1, 1, 2, 2, 3, 2, 1, 1, 2, 4, 2, 1, 3, 2, 2, 3, 3, 2, 1, 1, 1, 2, 0, 1, 2, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 2, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0
	7, 2, 2, 7, 2, 8, 7, 10, 9, 8, 20, 9, 9, 7, 7, 10, 11, 12, 14, 9, 9, 10, 14, 11, 9, 12, 13, 11, 12, 11, 9, 8, 7, 9, 9, 6, 9, 9, 7, 8, 7, 8, 9, 7, 6, 7, 7, 8, 6, 7, 6, 6, 2, 7, 7, 7, 6, 6, 7, 2, 6, 7, 6, 3, 6, 2, 6, 2, 2, 6, 2, 2, 2, 2, 2
wind	0, 2, 2, 4, 7, 5, 8, 5, 7, 5, 6, 5, 3, 3, 1, 1, 2, 1, 1, 0, 1
	2, 11, 11, 18, 28, 23, 27, 22, 26, 24, 25, 22, 16, 13, 9, 7, 11, 7, 8, 2, 6
rain	66, 1, 1, 1, 1, 0, 0
	238, 6, 3, 6, 3, 2, 3
area	32, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1
	113, 3, 2, 2, 2, 2, 2, 2, 6, 2, 6, 2, 2, 2, 6, 2, 2, 2, 2, 2, 3, 6, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 6, 3, 2, 6, 2, 3, 6, 6, 3, 2, 6, 2, 2, 6, 6, 2, 2, 2, 2, 2, 6, 2, 5, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 6, 2, 3, 2, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 6, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 2, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 6, 2, 2, 3, 6, 3, 2, 2, 2, 2, 2, 2, 2, 6, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 2, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 6, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 2, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3

Tabela 4. Czas wyszukiwania rekordów o określonej wartości w niektórych atrybutach ze zbioru OnlineNewsPopularity.

Nazwa atrybutu	**OnlineNewsPopularity**
min positive polarity	237, 2295, 1323, 679, 40, 3543, 6, 833, 20, 81, 98, 257, 23, 2, 119, 18, 29, 14, 17, 4, 1, 19, 25, 1, 4, 1, 48, 4, 2, 1, 2, 0, 1
	1327, 4878, 4189, 2255, 664, 10030, 540, 2891, 542, 756, 751, 1253, 537, 465, 888, 517, 545, 503, 507, 464, 511, 524, 530, 462, 472, 458, 602, 476, 467, 452, 461, 461, 492
max positive polarity	232, 2, 1, 1, 9, 13, 1, 3, 2, 15, 5, 1, 27, 5, 15, 9, 15, 5, 1, 30, 84, 5, 13, 3, 1216, 23, 744, 1, 11, 573, 6, 132, 996, 2, 124, 207, 1, 2695
	1280, 495, 469, 473, 493, 501, 463, 464, 469, 603, 470, 508, 547, 474, 499, 487, 500, 473, 512, 563, 705, 468, 489, 460, 4405, 674, 2938, 613, 511, 2301, 490, 924, 3826, 608, 915, 1149, 546, 9604
min negative polarity	935, 116, 75, 987, 123, 1, 99, 860, 164, 12, 89, 1275, 2, 7, 2289, 3, 48, 4, 1358, 49, 17, 1, 168, 15, 115, 648, 124, 442, 9, 2, 1, 8, 1, 527, 142, 5, 5, 354, 110, 31, 4, 21, 246, 122, 2, 2, 1, 29, 1, 43, 1, 5, 2, 863
	8445, 1487, 1160, 4794, 1923, 927, 1285, 4197, 1412, 861, 1195, 5559, 849, 922, 9651, 865, 1013, 822, 5845, 1079, 865, 822, 1472, 891, 1225, 3235, 1267, 2496, 816, 798, 839, 805, 807, 2664, 1203, 819, 893, 2021, 1209, 898, 845, 898, 1747, 1283, 825, 891, 799, 903, 788, 934, 840, 824, 784, 4040
max negative polarity	23, 3, 2, 15, 2, 2, 20, 5, 2, 33, 1, 131, 6, 184, 10, 3, 34, 4, 29, 158, 42, 172, 6, 1, 2, 8, 1, 458, 203, 10, 12, 581, 390, 155, 10, 82, 1778, 2940, 82, 58, 87, 868, 90, 3530, 85, 268, 89, 76, 898
	985, 859, 858, 914, 876, 870, 944, 871, 1020, 1486, 895, 1371, 924, 1494, 1272, 864, 927, 806, 939, 1407, 1195, 1458, 1795, 1001, 953, 849, 758, 2065, 1406, 770, 760, 2623, 2082, 1259, 802, 1064, 6590, 10519, 988, 917, 992, 3519, 989, 11689, 955, 1453, 958, 912, 3430

title subjectivity	24, 3, 3, 12, 0, 2, 1, 55, 1, 2, 8, 1, 129, 1, 2, 1, 3, 0, 11, 1, 1, 1, 1, 240, 2, 3, 2, 1, 60, 1, 7, 1, 5, 8, 38, 1, 2, 2, 2, 2, 1, 47, 1, 1, 6, 3, 2, 1, 5, 1, 1, 7, 5, 1, 1, 1, 1, 2, 1, 214, 1, 1, 1, 2, 1, 1, 2, 1, 4, 8, 9, 2, 2, 1, 3, 1, 6, 1, 31, 1, 1, 2, 1, 21, 1, 4, 1, 1, 1, 1, 1, 3, 1, 1, 1, 117, 1, 2, 1, 2, 1, 2, 1, 7, 2, 2, 3, 1, 12, 3, 2, 1, 3, 1, 1, 1, 7, 1, 1, 13, 3, 1, 1, 1, 2, 26, 1, 4, 3, 2, 1, 58, 2, 5, 2, 1, 1, 1, 283, 1, 2, 1, 1, 1, 1, 2, 2, 1, 4, 6, 2, 1, 1, 8, 2, 4, 1, 1, 1, 2, 3, 4, 6, 1, 8, 1, 2, 0, 2, 1, 1, 1, 1, 1, 190, 1, 3, 1, 1, 1, 1, 2, 1, 1, 2, 1, 1, 5, 1, 1, 1, 1, 2, 55, 1, 2, 1, 1, 3, 1, 4, 11, 1, 2, 1, 1, 1, 3, 2, 1, 1, 2, 1, 1, 1, 23, 1, 1, 1, 1, 1, 2, 4, 1, 122, 1, 11, 1, 2, 1, 1, 1, 15, 1, 2, 3, 4, 1, 1, 4, 1, 1, 2, 0, 2, 1, 1, 8, 6, 0, 2, 2, 1, 396, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 2, 2, 1, 1, 1, 2, 1, 1, 8, 25, 3, 1, 0, 2, 1, 1, 10, 1, 18, 1, 6, 2, 2, 2, 1, 1, 37, 2, 1, 1, 3, 1, 1, 13, 1, 1, 1, 1, 4, 1, 1, 11, 4, 1, 1, 2, 1, 1, 64, 1, 1, 1, 2, 1, 1, 1, 344, 4, 1, 1, 1, 1, 1, 1, 1, 5, 1, 1, 3, 2, 1, 1, 1, 14, 1, 1, 2, 1, 1, 1, 3, 1, 1, 2, 2, 1, 8, 1, 2, 27, 2, 1, 2, 0, 1, 1, 1, 1, 1, 9, 0, 5, 1, 1, 2, 2, 1, 1, 3, 3, 1, 2, 1, 2, 0, 1, 1, 1, 4, 3, 0, 772, 2, 1, 1, 1, 3, 2, 1, 1, 1, 3, 1, 3, 1, 1, 1, 1, 1, 7, 3, 2, 1, 1, 1, 1, 4, 2, 1, 2, 4, 1, 1, 13, 1, 1, 7, 1, 2, 1, 1, 1, 1, 1, 24, 1, 0, 23, 2, 1, 1, 1, 1, 1, 1, 2, 36, 2, 1, 1, 14, 1, 2, 2, 2, 1, 1, 2, 70, 1, 3, 1, 1, 4, 1, 2, 1, 3, 2, 5, 2, 1, 1, 1, 1, 1, 10, 1, 1, 1, 1, 1, 1, 1, 1, 16, 2, 1, 13, 5, 4, 1, 1, 1, 1, 1, 27, 1, 1, 2, 1, 1, 1, 1, 1, 1, 2, 1, 2, 2, 0, 1, 188, 3, 5, 2, 1, 2, 2, 2, 1, 1, 3, 1, 1, 1, 1, 9, 4, 1, 1, 1, 2, 2, 1, 1, 1, 33, 1, 5, 1, 15, 0, 1, 3, 1, 3, 41, 2, 6, 1, 1, 1, 1, 1, 99, 3, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 2, 1, 1, 1, 115, 2, 2, 11, 12, 1, 3, 1, 1, 1, 4, 2, 1, 17, 3, 1, 1, 2, 1, 1, 1, 1, 1, 161, 2, 4, 1, 3, 1, 1, 1, 1, 1, 5, 3, 2, 1, 2, 0, 2, 5, 29, 1, 10, 1, 1, 1, 1, 1, 159, 1, 1, 2, 2, 2, 1, 6, 2, 2, 5, 1, 1, 5, 1, 1, 1, 2, 3, 1, 2, 1, 3, 94, 1, 2, 2, 1, 2, 1, 1, 3, 1, 6, 48, 1, 41, 1, 5, 5, 1, 14, 1, 4, 1, 1, 185, 6, 3, 1, 1, 1, 1, 42, 1, 2, 745
	21872, 828, 762, 806, 761, 779, 786, 972, 765, 760, 786, 800, 1567, 822, 785, 1011, 773, 735, 760, 732, 725, 746, 739, 1547, 749, 797, 767, 755, 949, 715, 839, 746, 739, 744, 833, 773, 718, 717, 718, 703, 770, 877, 699, 706, 731, 726, 714, 707, 720, 687, 761, 735, 719, 715, 718, 769, 719, 718, 710, 1461, 689, 685, 706, 720, 789, 708, 718, 717, 713, 796, 725, 699, 738, 710, 771, 724, 735, 711, 816, 767, 707, 723, 699, 787, 790, 731, 711, 708, 706, 757, 721, 723, 708, 714, 810, 1108, 710, 727, 721, 723, 720, 719, 703, 746, 735, 720, 719, 704, 771, 717, 716, 707, 719, 723, 794, 710, 735, 713, 707, 826, 720, 718, 705, 706, 787, 804, 706, 703, 702, 775, 716, 917, 746, 771, 753, 733, 720, 705, 1700, 715, 723, 709, 716, 712, 716, 718, 715, 716, 733, 731, 709, 704, 718, 749, 723, 726, 711, 721, 722, 719, 720, 726, 748, 717, 739, 705, 713, 720, 721, 708, 704, 717, 874, 749, 1459, 752, 840, 725, 866, 928, 1116, 821, 770, 1045, 839, 739, 743, 754, 696, 743, 750, 727, 746, 918, 735, 725, 720, 673, 723, 732, 735, 745, 707, 720, 719, 709, 720, 720, 705, 717, 712, 726, 705, 716, 732, 797, 717, 708, 683, 738, 719, 707, 726, 706, 1143, 704, 753, 713, 711, 782, 708, 714, 763, 707, 757, 728, 728, 706, 707, 785, 721, 707, 719, 703, 790, 730, 710, 741, 707, 785, 718, 708, 705, 2168, 743, 746, 737, 819, 741, 719, 711, 706, 781, 697, 705, 721, 718, 808, 709, 661, 721, 719, 736, 725, 813, 727, 718, 800, 719, 717, 703, 745, 780, 779, 715, 730, 717, 799, 703, 706, 704, 837, 795, 706, 721, 724, 711, 804, 745, 715, 703, 706, 785, 733, 720, 834, 928, 905, 809, 1114, 975, 943, 1042, 975, 928, 810, 829, 745, 731, 807, 682, 1960, 811, 733, 725, 748, 738, 802, 755, 744, 787, 749, 842, 731, 716, 718, 691, 789, 763, 706, 708, 685, 795, 712, 714, 721, 712, 760, 712, 719, 706, 702, 800, 725, 800, 721, 707, 780, 711, 711, 673, 708, 787, 711, 746, 705, 728, 769, 701, 715, 720, 719, 798, 726, 719, 707, 693, 766, 746, 737, 729, 733, 779, 748, 748, 718, 3562, 729, 789, 721, 711, 723, 718, 775, 692, 707, 717, 719, 809, 709, 705, 716, 707, 791, 737, 722, 710, 700, 784, 713, 712, 741, 721, 755, 727, 724, 709, 708, 839, 699, 698, 733, 708, 787, 692, 709, 704, 709, 785, 804, 706, 718, 751, 740, 713, 719, 707, 706, 793, 710, 716, 829, 716, 758, 680, 758, 706, 716, 783, 727, 705, 704, 697, 1002, 720, 723, 710, 696, 800, 717, 725, 831, 778, 759, 936, 844, 941, 864, 1121, 883, 903, 925, 897, 855, 729, 821, 750, 726, 751, 714, 855, 745, 727, 778, 797, 745, 722, 735, 723, 786, 734, 813, 714, 706, 740, 700, 705, 709, 706, 731, 710, 733, 746, 740, 766, 731, 726, 1348, 740, 765, 730, 710, 725, 728, 735, 717, 712, 726, 708, 762, 720, 714, 741, 722, 779, 689, 705, 711, 719, 731, 707, 716, 847, 693, 751, 715, 767, 705, 710, 734, 708, 723, 854, 714, 760, 694, 719, 703, 698, 714, 1050, 736, 719, 706, 713, 716, 697, 720, 704, 707, 706, 709, 708, 713, 736, 704, 706, 1106, 727, 700, 741, 746, 693, 733, 717, 708, 707, 726, 716, 730, 765, 723, 683, 704, 737, 714, 706, 702, 707, 715, 1288, 682, 721, 734, 741, 717, 705, 712, 727, 723, 681, 720, 725, 732, 685, 703, 718, 727, 823, 711, 739, 703, 708, 726, 725, 708, 1262, 707, 728, 717, 717, 717, 720, 743, 679, 710, 723, 705, 714, 690, 698, 718, 713, 728, 844, 779, 771, 747, 979, 1096, 756, 877, 834, 854, 855, 833, 981, 791, 756, 750, 885, 739, 872, 721, 755, 770, 707, 778, 699, 728, 727, 709, 1364, 735, 733, 688, 705, 705, 710, 872, 713, 719, 3403

Powyżej zostały przedstawione czasy wyszukiwania rekordów o określonej wartości w konkretnych atrybutach. Pierwszym zbiorem poddanym analizie jest zbiór *ForestFires*. Posiada on 519 rekordów oraz 13 atrybutów. Drugim rozpatrywanym zbiorem jest dużo większy zbiór *OnlineNewsPopularity.* Składa się on z 39 644 rekordów oraz aż 60 atrybutów. Wyniki ze zbioru *RedWine* nie zostały tutaj przedstawione ze względu na fakt, iż wyniki czasowe są bardzo podobne jak te ze zbioru *ForestFires*. Mimo, że zbiór jest prawie trzykrotnie większy, to czasy wyszukiwania są na tyle małe, że prawdziwą rozbieżność widać dopiero przy porównaniu ze zbiorem tak dużym jak *OnlineNewsPopularity.*

Wiersze na tle brzoskwiniowym przedstawiają wyniki dotyczące grafów, natomiast te na tle niebieskim dotyczą tabeli.

Każda pojedyncza liczba przedstawia w mikrosekundach czas wyszukiwania wszystkich rekordów, które posiadają konkretną wartość, w konkretnym atrybucie. Jak widać w różnych atrybutach jest podana różna liczba czasów. Im więcej duplikatów wartości w atrybucie, tym mniej wartości czasowych w wierszu.

Patrząc na pierwsze wiersze dotyczące wyników ze zbioru *ForestFires* można zauważyć, że w tych atrybutach znajduje się mnóstwo duplikatów. Czasy wyszukiwań w takim przypadku dla operacji w grafie są odrobinę większe. Mimo, że listy z atrybutami mamy dostępne od ręki, to czas który jest liczony, mierzy również pobranie listy z rekordami, które są wynikiem. Przy bardzo dużych listach, w przypadku zbiorów, które zawierają mnóstwo duplikatów, ten czas jest niewiele większy. W przypadku zbioru *ForestFires* czasy operacji wyszukiwania w asocjacyjnym grafie są średnio **2,5-krotnie lepsze.**

Prawdziwą oszczędność czasu widać analizując zbiór *OnlineNewsPopularity*. Jest to duży zbiór danych, dlatego analizie poddano tylko kilka atrybutów. Przy blisko 40 tysiącach rekordów można zauważyć, jak ogromną oszczędnością jest pobieranie informacji z asocjacyjnego grafu. Prawdziwą różnicę widać w przypadku zbioru, gdzie występuje mniej duplikatów wartości. W atrybucie *„title subjectivity”*, gdzie znajduje się dużo więcej unikalnych wartości, niż w przypadku innych atrybutów, czasy operacji w grafach w porównaniu do tych wykonywanych w tabelach, są nawet **700-krotnie korzystniejsze.** Można sobie wyobrazić, ile czasu można oszczędzić przy jeszcze większych zbiorach danych!

3.4. Wyszukiwanie relacji koniunkcji i alternatywy

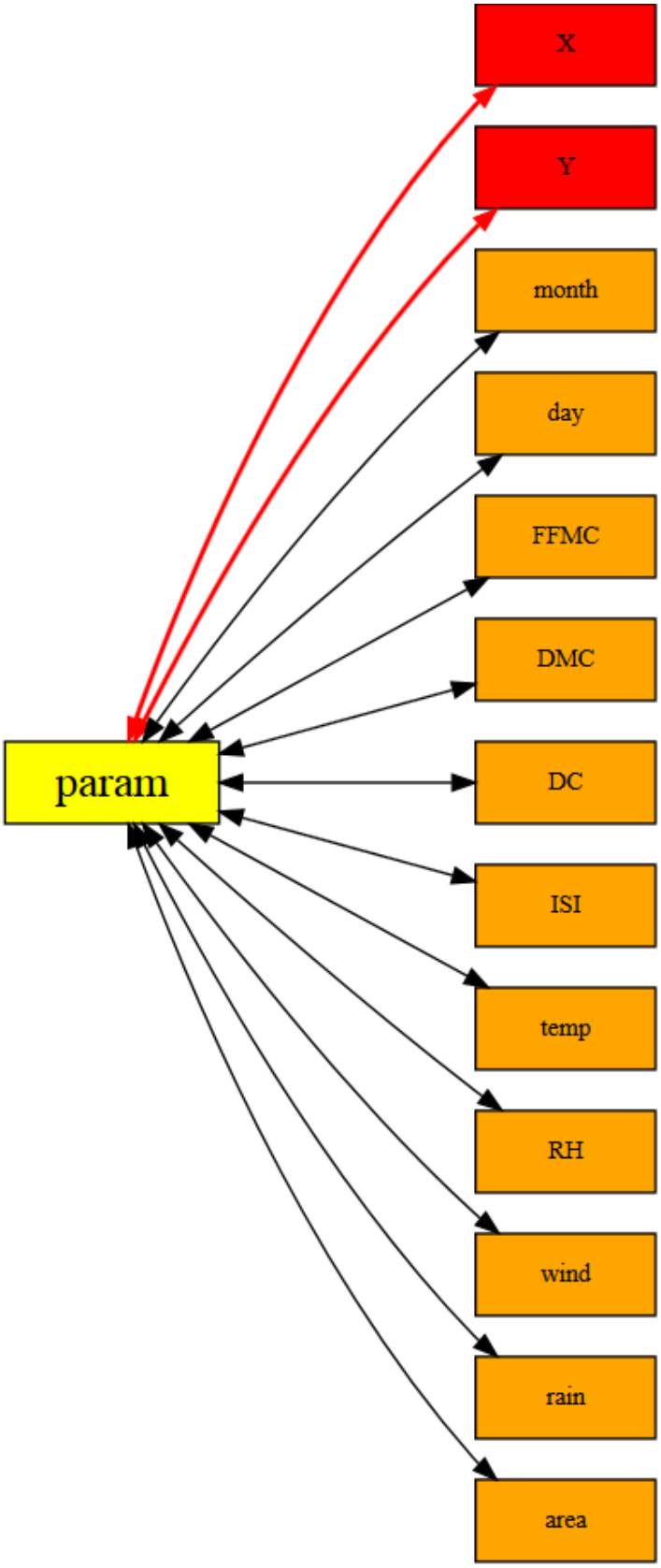
3.4.1. Wyszukiwanie relacji koniunkcji i alternatywy w grafie

Kolejnymi relacjami, które w bardzo efektywny sposób jesteśmy w stanie znaleźć w grafie, są relacje koniunkcji i alternatywy. W rozbudowanych zbiorach danych szybkie wyszukiwanie tego typu informacji jest niezwykle istotne. Dzięki zaimplementowanemu rozwiązaniu jesteśmy w stanie znaleźć wynik koniunkcji oraz alternatywy dowolnej liczby zmiennych.

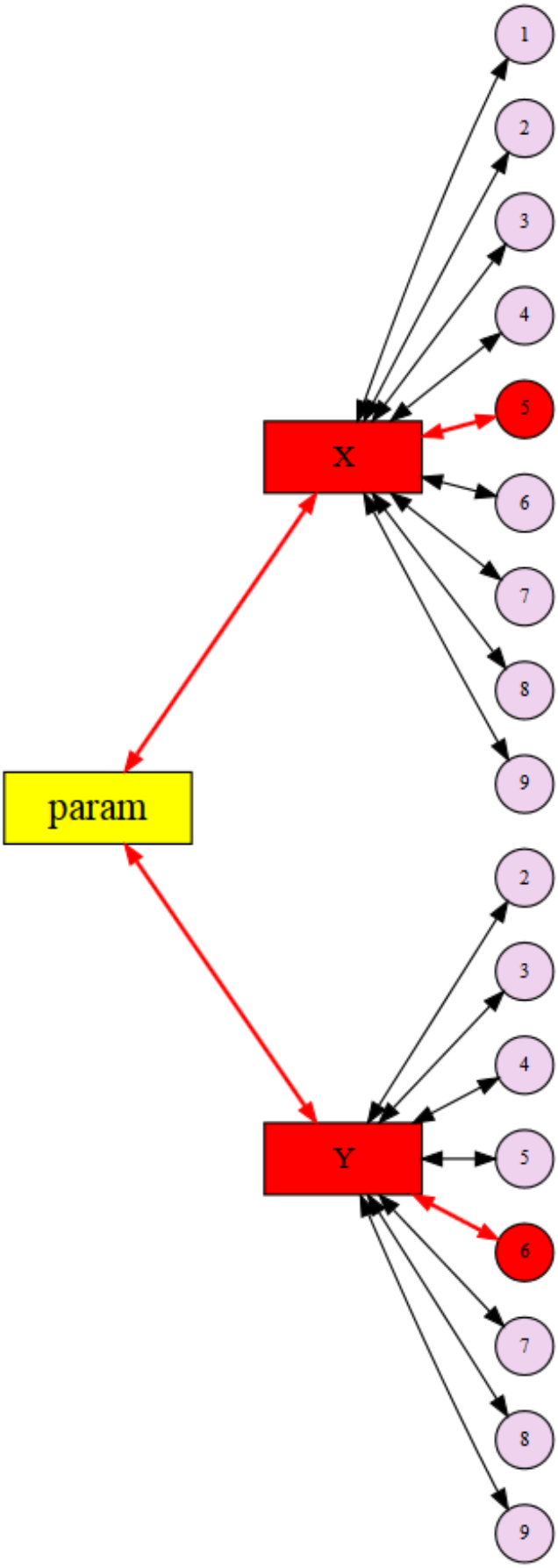
W proponowanym rozwiązaniu jedyne, co użytkownik powinien podać, to szukane wartości oraz ich atrybuty. Następnie znajdywane są wszystkie rekordy, które reprezentowane są przez podaną wartość w konkretnym atrybucie. Uściślić trzeba, że rekordy te są nie tyle znajdywane, co dostępne od ręki, dzięki implementacji asocjacyjnego grafu. W tym momencie wszystkie wskazane rekordy są pobudzane wartością 1. Dzieje się tak ze wszystkimi wskazanymi rekordami. Zdarza się, że rekordy są wielokrotnie pobudzane, jeżeli wskazana wartość dotyczy ich w kilku atrybutach. Dzięki operacji pobudzania jedynką, dostajemy informację o tym, ile z szukanych wartości tak naprawdę dotyczy poszczególnych rekordów.

Złożoność obliczeniowa takiego algorytmu wynosi O(n), co oznacza, że w zależności od ilości znalezionych rekordów należy wykonać maksymalnie n operacji dominujących. Operacją dominującą jest tutaj pobudzenie rekordu wartością 1. Liczba iteracji n będzie mniejsza lub równa ilości wszystkich rekordów. Jednak w większości zbiorów danych, n będzie liczbą kilkukrotnie mniejszą, niż liczba wszystkich rekordów.

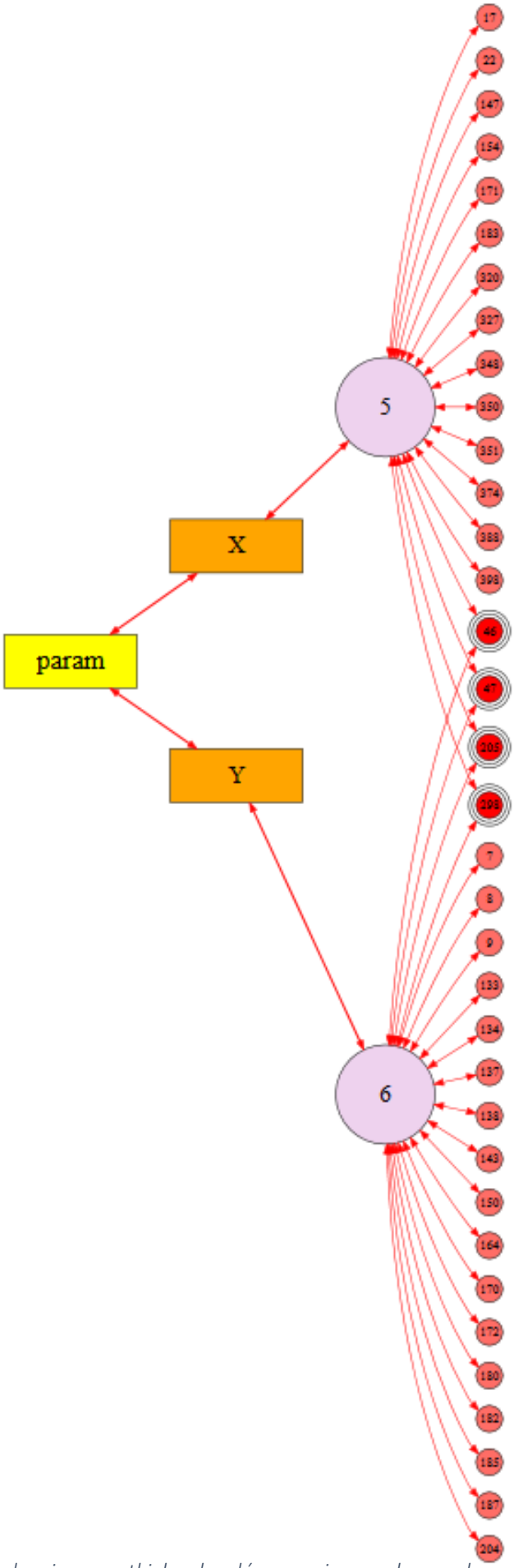
Kolejne etapy wyszukiwania koniunkcji i alternatywy w grafie



Rysunek 8. Wskazanie atrybutów, których dotyczy wyszukiwanie relacji koniunkcji i alternatywy.



Rysunek 9. Wskazanie wartości, dla których będą wyszukiwane relacje koniunkcji i alternatywy.



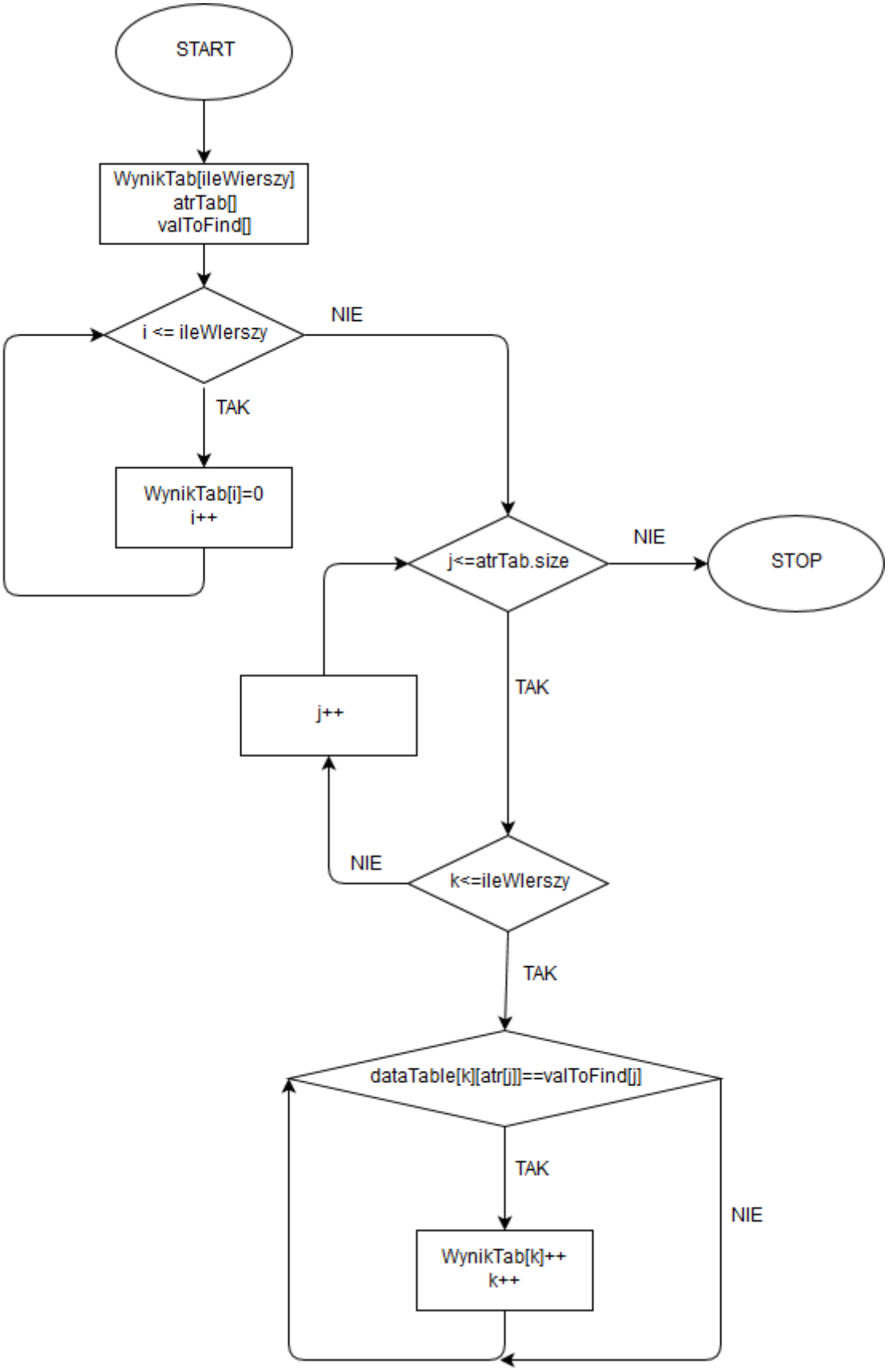
Rysunek 10. Pobudzenie wszystkich rekordów, powiązanych ze wskazanymi wartościami.

3.4.2. Wyszukiwanie relacji koniunkcji i alternatywy w tabeli

Wyszukiwanie relacji koniunkcji i alternatywy w tabeli jest bardziej skomplikowane, gdy zbiór danych zagregowany jest w tabeli.

Aby znaleźć wynik relacji koniunkcji i alternatywy należy najpierw stworzyć kontener, w którym będą przechowywane wartości ilości pobudzeń. Należy wykonać tak naprawdę kilka iteracji i to po całym zbiorze danych. Jest to ogromna ilość iteracji w dużych zbiorach danych, takich jak na przykład *OnlineNewsPopularity*. Pierwsza iteracja ma na celu inicjalizację tablicy wynikowej zerami. Następnie, w zależności od wybranej ilości wartości do wykonania operacji koniunkcji i alternatywy (stała liczba dla operacji konkretnej ilości wartości), tyle razy powtarzana jest znów iteracja po wszystkich rekordach ze zbioru. W każdej pojedynczej operacji sprawdzany jest warunek, czy szukana wartość jest równa wartości z konkretnego atrybutu i rekordu. Jeżeli tak, w tabeli wynikowej wartość powtórzeń zostaje powiększona o jeden dla tego rekordu.

Złożoność obliczeniowa takiego algorytmu po przeliczeniu również wynosi O(n). Biorąc pod uwagę, że ilość wartości szukanych w relacji koniunkcji i alternatywy jest stała dla konkretnego przypadku, wtedy otrzymujemy taki wynik. Operacja dominująca składa się z dwóch podoperacji – porównania wartości oraz, jeśli wynik porównania jest prawdziwy, powiększenia wyniku powtórzeń dla konkretnego rekordu.



Rysunek 11. Algorytm relacji koniunkcji i alternatywy w tabeli.

3.4.3. Przykłady operacji koniunkcji i alternatywy

Jako pierwszy zostanie przedstawiony przykład ze zbioru *ForestFires*, zobrazowany wcześniej na grafach.

Jaką informację chcemy uzyskać?

- Listę rekordów reprezentującą pożary lasów na terenach oznaczonych współrzędną przestrzenną osi X w Montesinho Park (oznaczenia 1 – 9) równą 5 lub współrzędną przestrzenna osi Y (oznaczenia 2 – 9) równą 6.
- Listę rekordów reprezentującą pożary lasów na terenie wyznaczonym dokładnie przez współrzędne [X,Y] = [5,6].

Jest to dokładnie ta sama operacja, która została przedstawiona wcześniej na grafach. Jako wynik zostanie przedstawiona lista wszystkich rekordów z uzyskaną liczbą powtórzeń. Przy dwóch wartościach branych jako dane wejściowe do wyliczenia relacji koniunkcji i alternatywy, wynik będzie interpretowany w poniższy sposób:

- 2 – relacja koniunkcji. Oznacza to, że rekord, który uzyskał taką wartość, reprezentuje pożar na terenie oznaczonym dokładnie przez współrzędne [X,Y]=[5,6].
- 1 – relacja alternatywy. Rekord, który uzyskał taką wartość, reprezentuje pożar na terenie oznaczonym współrzędną X =5 lub Y=6.
- 0 – rekord, który uzyskał taką wartość, reprezentuje pożar na terenie oznaczonym przez współrzędną X różną od 5 oraz współrzędną Y różną od 6.

Dodatkowo zostanie podany czas w mikrosekundach znalezienia wyniku w grafie oraz w tabeli.

Tabela 5. Wyniki operacji koniunkcji i alternatywy dla dwóch wartości i zbioru ForestFires.

Numer rekordu i wartość powtórzeń											
398	2	462	1	322	0	475	0	33	0	279	0
46	2	487	1	505	0	406	0	32	0	278	0
47	2	488	1	508	0	405	0	31	0	277	0
205	2	459	1	319	0	472	0	30	0	273	0
384	1	9	1	318	0	473	0	29	0	271	0
137	1	73	1	317	0	474	0	28	0	270	0
138	1	75	1	316	0	403	0	27	0	269	0
404	1	8	1	509	0	402	0	26	0	268	0
147	1	74	1	315	0	401	0	25	0	267	0
150	1	7	1	314	0	259	0	24	0	266	0
154	1	6	1	313	0	400	0	38	0	214	0
392	1	5	1	312	0	399	0	21	0	264	0
388	1	4	1	466	0	494	0	20	0	263	0
134	1	345	0	428	0	476	0	19	0	262	0
164	1	368	0	463	0	397	0	18	0	261	0
383	1	365	0	426	0	396	0	16	0	260	0
382	1	500	0	464	0	477	0	15	0	1	0
381	1	343	0	424	0	395	0	14	0	258	0
170	1	366	0	465	0	394	0	13	0	256	0
171	1	341	0	422	0	393	0	12	0	255	0
433	1	340	0	421	0	478	0	11	0	254	0
446	1	502	0	420	0	479	0	10	0	253	0
442	1	338	0	419	0	391	0	3	0	252	0
92	1	503	0	418	0	390	0	2	0	251	0
93	1	337	0	417	0	87	0	55	0	159	0
94	1	336	0	429	0	105	0	69	0	177	0
95	1	335	0	467	0	104	0	68	0	176	0
441	1	367	0	414	0	103	0	67	0	175	0
438	1	334	0	413	0	102	0	66	0	174	0
437	1	333	0	412	0	101	0	65	0	173	0
377	1	332	0	411	0	100	0	64	0	169	0
427	1	504	0	410	0	99	0	63	0	168	0
425	1	326	0	409	0	98	0	62	0	167	0
423	1	360	0	408	0	97	0	61	0	166	0
416	1	359	0	407	0	96	0	60	0	165	0
415	1	358	0	468	0	91	0	59	0	163	0
131	1	357	0	469	0	90	0	58	0	162	0
132	1	356	0	457	0	89	0	57	0	161	0
133	1	355	0	454	0	88	0	56	0	160	0
295	1	361	0	453	0	106	0	141	0	178	0
236	1	354	0	452	0	86	0	54	0	158	0
320	1	353	0	451	0	85	0	53	0	157	0
247	1	346	0	450	0	84	0	52	0	156	0

249	1	362	0	449	0	83	0	51	0	155	0
257	1	363	0	448	0	82	0	50	0	153	0
304	1	364	0	447	0	81	0	49	0	152	0
303	1	352	0	456	0	80	0	45	0	151	0
299	1	497	0	445	0	79	0	44	0	149	0
298	1	498	0	444	0	78	0	43	0	148	0
327	1	349	0	443	0	77	0	42	0	146	0
272	1	499	0	470	0	76	0	41	0	145	0
294	1	347	0	458	0	72	0	40	0	144	0
274	1	514	0	440	0	71	0	39	0	142	0
275	1	310	0	439	0	120	0	232	0	197	0
276	1	309	0	460	0	140	0	248	0	213	0
281	1	308	0	461	0	139	0	246	0	212	0
287	1	307	0	436	0	136	0	245	0	211	0
285	1	306	0	435	0	135	0	244	0	210	0
350	1	305	0	434	0	130	0	243	0	209	0
143	1	511	0	432	0	129	0	242	0	208	0
374	1	302	0	431	0	128	0	241	0	207	0
180	1	301	0	430	0	127	0	240	0	206	0
182	1	300	0	376	0	126	0	239	0	203	0
183	1	513	0	480	0	125	0	238	0	202	0
185	1	311	0	387	0	124	0	237	0	201	0
187	1	297	0	386	0	123	0	235	0	200	0
204	1	296	0	385	0	122	0	234	0	199	0
351	1	515	0	481	0	121	0	233	0	198	0
172	1	293	0	482	0	70	0	250	0	286	0
348	1	292	0	483	0	119	0	231	0	196	0
344	1	291	0	380	0	118	0	227	0	195	0
342	1	290	0	379	0	117	0	226	0	194	0
339	1	289	0	485	0	116	0	225	0	193	0
221	1	288	0	486	0	115	0	224	0	192	0
228	1	516	0	378	0	114	0	223	0	191	0
229	1	517	0	389	0	113	0	222	0	190	0
230	1	321	0	375	0	112	0	220	0	189	0
495	1	331	0	489	0	111	0	219	0	188	0
22	1	330	0	373	0	110	0	218	0	186	0
471	1	329	0	372	0	109	0	217	0	184	0
501	1	328	0	490	0	108	0	216	0	181	0
17	1	506	0	371	0	107	0	215	0	179	0
48	1	507	0	455	0	23	0	265	0		
510	1	491	0	492	0	37	0	284	0		
496	1	325	0	370	0	36	0	283	0		
484	1	324	0	493	0	35	0	282	0		
512	1	323	0	369	0	34	0	280	0		

[GRAF] CZAS W MIKROSEKUNDACH 69
[TABELA] CZAS W MIKROSEKUNDACH 237

Inne przykłady operacji koniunkcji i alternatywy

Jakie informacje chcemy uzyskać?

- Listę rekordów reprezentującą pożary lasów, które zdarzyły się dokładnie w lipcu, w czwartek, przy temperaturze 30.2 stopni Celsjusza
- Listę rekordów reprezentującą pożary lasów, które zdarzyły się w lipcu lub zdarzyły się w czwartek lub doszło do nich przy temperaturze powietrza 30.2 stopni Celsjusza.

Tabela 6. Wyniki operacji koniunkcji i alternatywy dla trzech wartości i zbioru ForestFires.

Numer rekordu i wartość powtórzeń											
482	3	107	1	374	1	139	1	476	1	254	1
481	3	76	1	170	1	134	1	478	1	255	1
477	2	324	1	166	1	402	1	479	1	256	1
325	1	323	1	372	1	404	1	506	1	286	1
426	1	454	1	379	1	192	1	288	1	475	1
427	1	455	1	162	1	410	1	289	1	276	1
429	1	456	1	381	1	197	1	290	1	48	1
431	1	457	1	382	1	198	1	41	1	469	1
210	1	322	1	384	1	413	1	291	1	55	1
433	1	321	1	177	1	120	1	317	1	51	1
436	1	320	1	153	1	416	1	292	1	56	1
88	1	66	1	152	1	417	1	293	1		
87	1	319	1	388	1	423	1	294	1		
86	1	318	1	150	1	206	1	295	1		
444	1	63	1	146	1	480	1	29	1		
224	1	144	1	85	1	287	1	492	1		
[GRAF] CZAS W MIKROSEKUNDACH 98 [TABELA] CZAS W MIKROSEKUNDACH 225											

Powyżej przedstawiono wyniki dla wyszukiwania relacji koniunkcji i alternatywy dla trzech wartości z różnych atrybutów dla zbioru *ForestFires*. Dla oszczędności miejsca przedstawiono wyniki jedynie z niezerowymi wartościami. Jak widać istnieją dwa rekordy w całym zbiorze danych, które zawierają koniunkcję trzech wskazanych wartości, jeden rekord, dla którego występują dwie z trzech wartości oraz osiemdziesiąt osiem rekordów z jedną wartością. Wszystkie wskazane rekordy są pozytywnym wynikiem dla operacji alternatywy.

Jaką informacje chcemy uzyskać?

- Listę rekordów zawierającą informacje o czerwonych winach, w których ilość całkowitego dwutlenku siarki wynosi 13, jakość jest oceniana na 8 (w skali od 0 do 10), pH posiada wartość 3.23 i zawartość alkoholu wynosi 12.6 procenta.

- Listę rekordów zawierającą informacje o czerwonych winach, w których ilość całkowitego dwutlenku siarki wynosi 13 lub jakość jest oceniana na 8 (w skali od 0 do 10) lub pH posiada wartość 3.23 lub zawartość alkoholu wynosi 12.6 procenta.

Tabela 7. Wyniki operacji koniunkcji i alternatywy dla czterech wartości i zbioru RedWine

Numer rekordu i wartość powtórzeń										
279	4	268	1	1270	1	973	1	1017	1	1224	1
1404	3	1489	1	1121	1	1254	1	444	1	123	1
1450	2	1150	1	664	1	975	1	792	1	257	1
672	1	1109	1	91	1	1062	1	1009	1	1481	1
1138	1	1234	1	499	1	1137	1	538	1	456	1
1336	1	328	1	911	1	142	1	802	1	1091	1
674	1	1550	1	1125	1	1462	1	1203	1	817	1
1492	1	1154	1	496	1	843	1	441	1	589	1
191	1	1347	1	482	1	794	1	829	1	1218	1
332	1	1451	1	508	1	137	1	1472	1	254	1
874	1	1441	1	660	1	590	1	1013	1	1027	1
761	1	726	1	562	1	525	1	38	1	1214	1
331	1	906	1	1499	1	446	1	391	1	1039	1
[GRAF] CZAS W MIKROSEKUNDACH 60 **[TABELA] CZAS W MIKROSEKUNDACH 304**										

Jaką informacje chcemy uzyskać?

- Listę rekordów zawierającą informacje o publikowanych w sieci artykułach, gdzie liczba słów w tytule wynosi 5, liczba linków 15, ilość zdjęć umieszczonych w artykule to 20, artykuł został opublikowany w sobotę oraz wskaźnik słów pozytywnych wynosi 1.

- Listę rekordów zawierającą informacje o publikowanych w sieci artykułach, gdzie liczba słów w tytule wynosi 5 lub liczba linków 15 lub ilość zdjęć umieszczonych w artykule to 20 lub artykuł został opublikowany w sobotę lub wskaźnik słów pozytywnych wynosi 1.

Tabela 8. Wyniki operacji koniunkcji i alternatywy dla pięciu wartości i zbioru OnlineNewsPopularity.

Numer rekordu i wartość powtórzeń											
6019	3	7749	2	13750	2	21222	2	26997	2	33117	2
8473	3	8130	2	13751	2	21560	2	27002	2	33520	2
322	2	8816	2	14799	2	21562	2	27350	2	33534	2
324	2	9171	2	14817	2	21891	2	27459	2	33690	2
330	2	9427	2	15174	2	21908	2	27710	2	33979	2
331	2	9478	2	15342	2	22144	2	27865	2	33990	2
716	2	9689	2	15535	2	22248	2	27868	2	33993	2
1128	2	9872	2	15544	2	22616	2	27909	2	34047	2
1396	2	10229	2	16199	2	23012	2	28146	2	34483	2
1969	2	10233	2	16200	2	23026	2	28312	2	34484	2
2065	2	10234	2	16209	2	23113	2	28327	2	34485	2
2382	2	10250	2	16212	2	23410	2	28775	2	34494	2
2733	2	10586	2	16818	2	23416	2	29140	2	34949	2
3522	2	10594	2	16822	2	23417	2	29221	2	34950	2
3531	2	10595	2	17177	2	23419	2	29657	2	34964	2
3916	2	10600	2	17538	2	23572	2	29676	2	35679	2
3918	2	10664	2	17540	2	23810	2	30534	2	35925	2
4273	2	10950	2	17891	2	24185	2	30926	2	35936	2
4620	2	11641	2	17902	2	24198	2	30928	2	36414	2
4946	2	12022	2	17908	2	24601	2	30936	2	36900	2
5659	2	12024	2	18687	2	24606	2	30940	2	36909	2
5672	2	12027	2	19346	2	24608	2	31319	2	36919	2
6025	2	12348	2	19351	2	25368	2	31321	2	37413	2
6031	2	12642	2	19362	2	25371	2	31763	2	37773	2
6365	2	12653	2	19726	2	25708	2	31785	2	37794	2
6375	2	12659	2	20090	2	25762	2	31790	2	38236	2
6712	2	12771	2	20096	2	25764	2	32195	2	38259	2
6718	2	12998	2	20112	2	25765	2	32196	2	38721	2
6727	2	13360	2	20640	2	25767	2	32202	2	39161	2
7050	2	13362	2	20665	2	26128	2	32205	2	39178	2
7068	2	13367	2	20846	2	26137	2	32762	2	5	1
7698	2	13368	2	20853	2	26138	2	32932	2	7	1
7733	2	13379	2	20862	2	26554	2	33112	2	…	1
[GRAF] CZAS W MIKROSEKUNDACH 2941 **[TABELA] CZAS W MIKROSEKUNDACH 7695**											

Powyżej została przedstawiona tabela z wynikami, ale tylko dla trzech i dwóch znalezionych wartości. Liczba rekordów, które zostały raz pobudzone wynosi 5052 rekordy. Aby nie zaciemniać wyniku nie zostały przedstawione tutaj numery tych rekordów, jednakże przedstawiony czas obejmuje znalezienie wszystkich wartości.

3.4.4. Wnioski dla wyszukiwania relacji koniunkcji i alternatywy

Jak można zauważyć wyszukiwanie relacji koniunkcji i alternatywy jest operacją dającą dużo informacji o rekordach znajdujących się w zbiorze. Informacje, które uzyskujemy, to coś więcej niż tylko wiadomość o koniunkcji i alternatywie, ale również dokładna informacja o liczbie pobudzeń danego rekordu, a więc sile danej alternatywy.

Konkretne informacje zostały wyszukane w trzech zbiorach o różniej wielkości oraz o różniej liczbie atrybutów.

- *ForestFires* zawierający 13 atrybutów oraz 517 rekordów
    - wyszukiwanie dwóch wartości:

      Czas wyszukiwania w grafie 3.4 razy korzystniejszy niż w tabeli
- *RedWine* zawierający 12 atrybutów oraz 1599 rekordów
    - wyszukiwanie trzech wartości:

      Czas wyszukiwania w grafie 2.3 razy korzystniejszy niż w tabeli
    - wyszukiwanie czterech wartości:

      Czas wyszukiwania w grafie 5 razy korzystniejszy niż w tabeli
- *OnlineNewsPopularity* zawierający 60 atrybutów oraz 39 644 rekordy
    - wyszukiwanie pięciu wartości:

      Czas wyszukiwania w grafie 2.6 razy korzystniejszy niż w tabeli

3.5. Obliczanie podobieństwa względem atrybutu

3.5.1. Obliczanie podobieństwa względem atrybutu w grafie

Kolejną bardzo ważną funkcjonalnością, którą dzięki konstrukcji grafu jesteśmy w stanie w bardzo szybki sposób policzyć, jest podobieństwo rekordów względem jednego atrybutu i konkretnej wartości. Oczywiście może to być wartość zupełnie z zewnątrz lub wartość definiująca konkretny rekord. W sytuacji, kiedy jest to wartość konkretnego rekordu, wtedy zostanie policzone podobieństwo wszystkich rekordów względem niego.

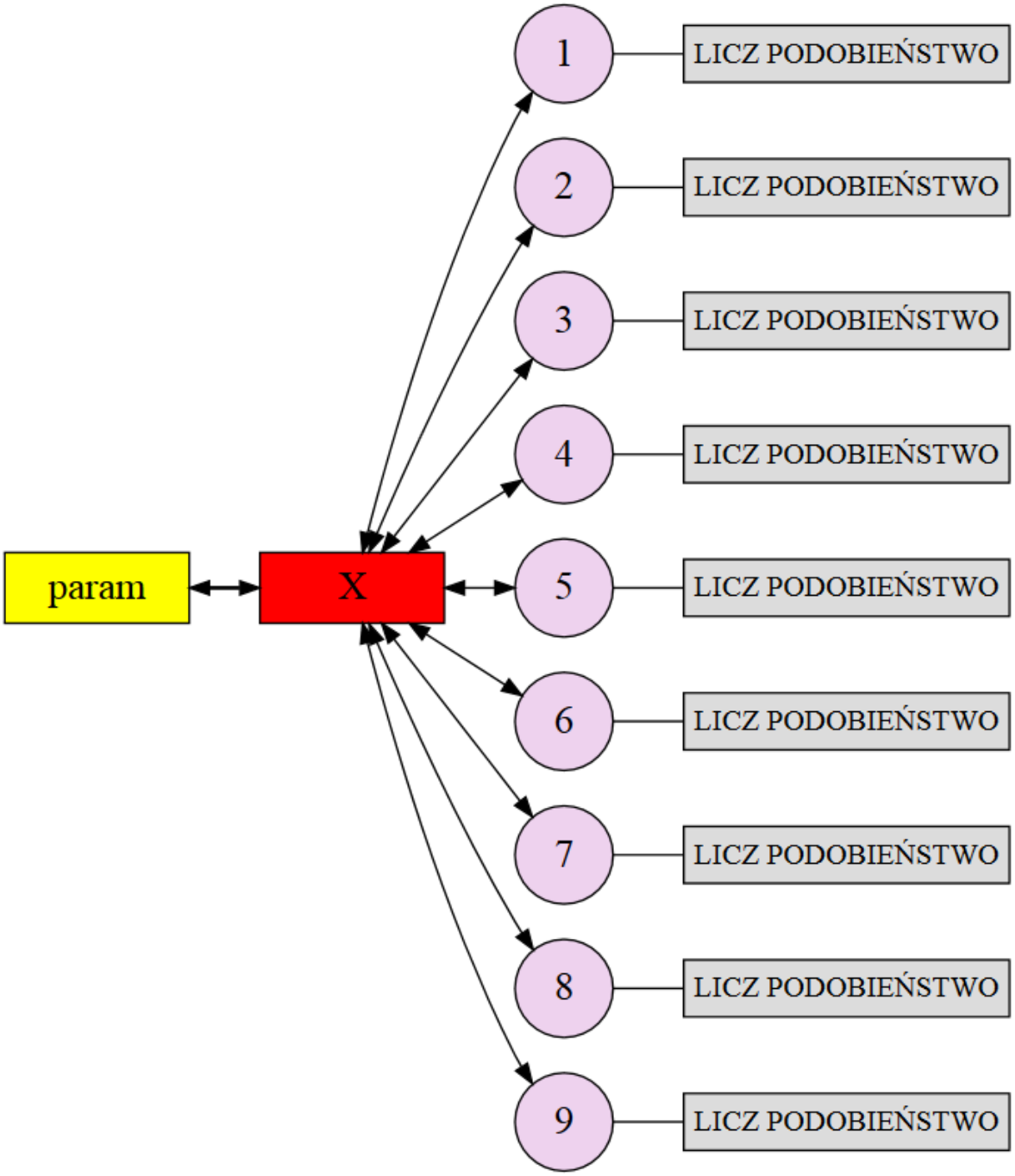
Wzór, którym jest liczone podobieństwo:

$$\mathrm{P} = 1 - \frac{|Val_A - Val_B|}{MAX - MIN}$$

Równanie 1. Wzór na podobieństwo.

Złożoność obliczeniowa takiego algorytmu wynosi O(n), co oznacza, że w zależności od ilości niezduplikowanych wartości należy wykonać n operacji dominujących. Operacją dominującą jest obliczenie podobieństwa.

Należy zaznaczyć, że ilość niezduplikowanych wartości reprezentujących atrybut w rzeczywistości jest kilkukrotnie mniejsza niż ilość wszystkich rekordów, co daje ogromną oszczędność obliczeniową w stosunku do obliczeń wykonywanych w tabeli.



Rysunek 12. Fragment asocjacyjnego grafu dla zbioru ForestFires przedstawiający atrybut reprezentujący współrzędną przestrzenną osi X oraz jego posortowany zbiór obiektów wartości wraz z zaznaczonymi miejscami wykonywania operacji liczenia podobieństwa.

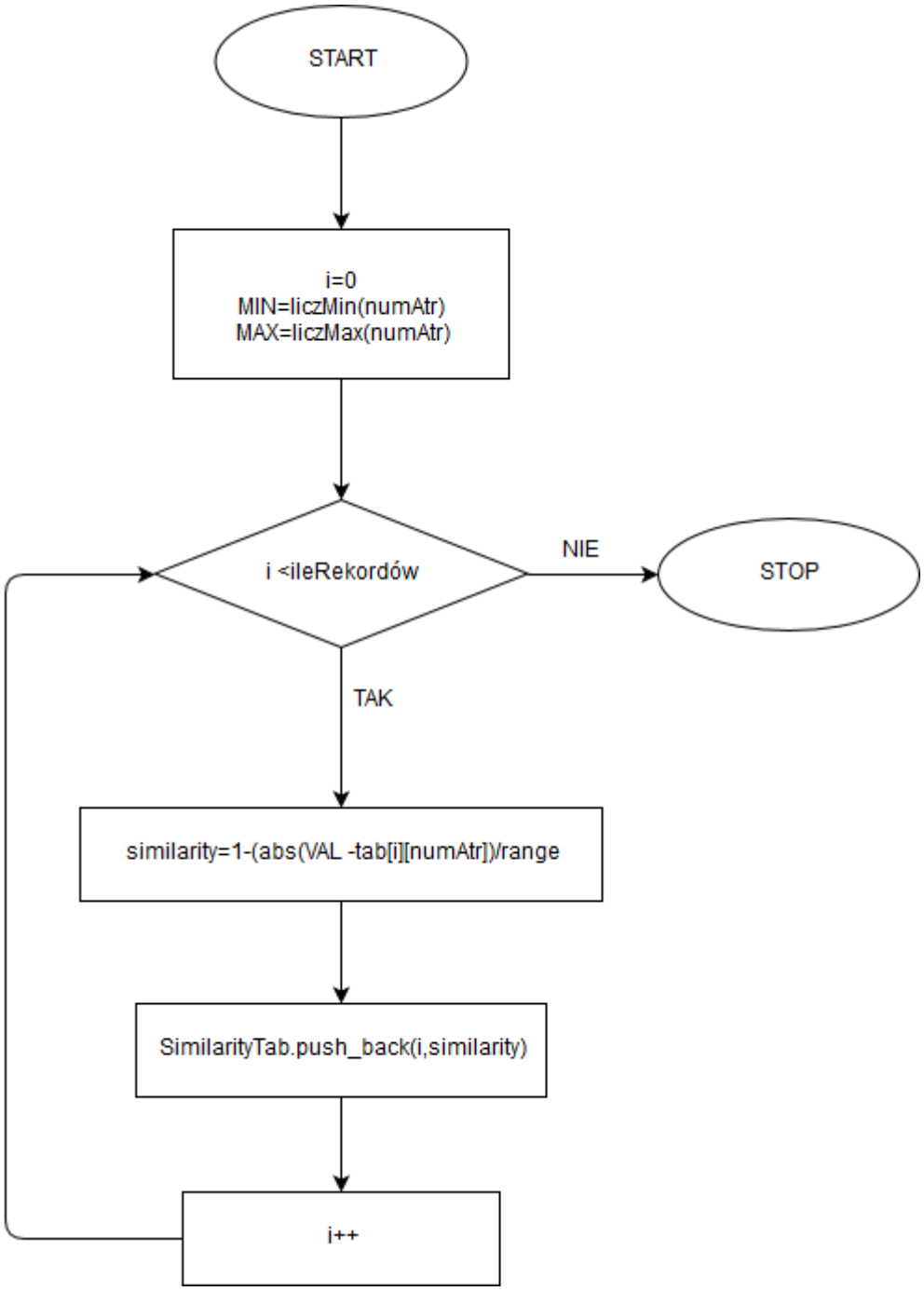
3.5.2. Obliczanie podobieństwa względem atrybutu w tabeli

Obliczanie podobieństwa względem atrybutu w tabeli jest niezwykle kosztowne, szczególnie przy dużych zbiorach danych, gdzie jest mnóstwo zduplikowanych wartości w konkretnym atrybucie.

Tak naprawdę, obliczenie podobieństwa dla wszystkich rekordów ze zbioru wymaga żmudnej iteracji po wszystkich elementach zbioru. Sytuacja staje się skomplikowana w momencie, gdy bierzemy pod uwagę bardzo duży zbiór danych z dużą ilością duplikatów. Wtedy tak naprawdę widać jak bardzo korzystnym jest skorzystanie ze struktury grafowej AGDS i tam wyliczenie wartości podobieństwa.

Złożoność obliczeniowa takiego algorytmu wynosi O(n), co oznacza, że w zależności od ilości wszystkich rekordów ze zbioru należy wykonać n operacji dominujących. Na operację dominującą składa się obliczenie podobieństwa oraz zapisanie go do tablicy wyników.

Mimo, że złożoność obliczeniowa operacji w grafie, jak i w tabeli jest taka sama, to wyliczanie podobieństwa w grafie jest dużo bardziej korzystne, niż obliczanie podobieństwa w tabeli. W grafie ilość iteracji zależy od ilości niezduplikowanych wartości w grafie. W tabeli zaś należy wykonać tyle iteracji, ile jest rekordów w zbiorze.



Rysunek 13. Algorytm obliczania podobieństwa dla wszystkich rekordów tabeli.

3.5.3. Wyniki czasowe dla obliczenia podobieństwa

Tabela 9. Wyniki czasowe obliczania podobieństwa w trzech zbiorach różniej wielkości.

Wartość i obliczone podobieństwo											
ATRYBUT TEMP											
Val	**Sim**	**Val**	**Sim**	**Val**	**Sim**	**Val**	**Sim**	**Val**	**Sim**	**Val**	**Sim**
2.2	0.267	11.6	0.569	15.7	0.701	19.1	0.810	22.5	0.920	26.2	0.961
4.2	0.331	11.7	0.572	15.8	0.704	19.2	0.814	22.6	0.923	26.3	0.958
4.6	0.344	11.8	0.576	15.9	0.707	19.3	0.817	22.7	0.926	26.4	0.955
4.8	0.350	12.2	0.588	16	0.711	19.4	0.820	22.8	0.929	26.7	0.945
5.1	0.360	12.3	0.592	16.1	0.714	19.5	0.823	22.9	0.932	26.8	0.942
5.2	0.363	12.4	0.595	16.2	0.717	19.6	0.826	23	0.936	26.9	0.939
5.3	0.367	12.6	0.601	16.3	0.720	19.7	0.830	23.1	0.939	27.2	0.929
5.5	0.373	12.7	0.605	16.4	0.723	19.8	0.833	23.2	0.942	27.3	0.926
5.8	0.383	12.8	0.608	16.6	0.730	19.9	0.836	23.3	0.945	27.4	0.923
6.7	0.412	12.9	0.611	16.7	0.733	20.1	0.842	23.4	0.949	27.5	0.920
7.5	0.437	13.1	0.617	16.8	0.736	20.2	0.846	23.5	0.952	27.6	0.916
8	0.453	13.2	0.621	16.9	0.740	20.3	0.849	23.6	0.955	27.7	0.913
8.2	0.460	13.3	0.624	17	0.743	20.4	0.852	23.7	0.958	27.8	0.910
8.3	0.463	13.4	0.627	17.1	0.746	20.5	0.855	23.8	0.961	27.9	0.907
8.7	0.476	13.7	0.637	17.2	0.749	20.6	0.859	23.9	0.965	28	0.904
8.8	0.479	13.8	0.640	17.3	0.752	20.7	0.862	24	0.968	28.2	0.897
8.9	0.482	13.9	0.643	17.4	0.756	20.8	0.865	24.1	0.971	28.3	0.894
9	0.486	14	0.646	17.6	0.762	20.9	0.868	24.2	0.974	28.6	0.884
9.3	0.495	14.1	0.650	17.7	0.765	21	0.871	24.3	0.977	28.7	0.881
9.8	0.511	14.2	0.653	17.8	0.768	21.1	0.875	24.5	0.984	28.9	0.875
10.1	0.521	14.3	0.656	17.9	0.772	21.2	0.878	24.6	0.987	29.2	0.865
10.2	0.524	14.4	0.659	18	0.775	21.3	0.881	24.8	0.994	29.3	0.862
10.3	0.527	14.5	0.662	18.1	0.778	21.4	0.884	24.9	0.997	29.6	0.852
10.4	0.531	14.6	0.666	18.2	0.781	21.5	0.887	**25**	**1**	30.2	0.833
10.5	0.534	14.7	0.669	18.3	0.785	21.6	0.891	25.1	0.997	30.6	0.820
10.6	0.537	14.8	0.672	18.4	0.788	21.7	0.894	25.3	0.990	30.8	0.814
10.9	0.547	14.9	0.675	18.5	0.791	21.8	0.897	25.4	0.987	31	0.807
11	0.550	15.1	0.682	18.6	0.794	21.9	0.900	25.5	0.984	32.3	0.765
11.2	0.556	15.2	0.685	18.7	0.797	22.1	0.907	25.6	0.981	32.4	0.762
11.3	0.559	15.4	0.691	18.8	0.801	22.2	0.910	25.7	0.977	32.6	0.756
11.4	0.563	15.5	0.695	18.9	0.804	22.3	0.913	25.9	0.971	33.1	0.740
11.5	0.566	15.6	0.698	19	0.807	22.4	0.916	26.1	0.965	33.3	0.733
[GRAF] CZAS W MIKROSEKUNDACH 72											
[TABELA] CZAS W MIKROSEKUNDACH 166											
ATRYBUT WIND											
0.4	0.600	2.2	0.800	**4**	**1**	5.8	0.800	7.6	0.600	9.4	0.400
0.9	0.656	2.7	0.856	4.5	0.944	6.3	0.744	8	0.556		
1.3	0.700	3.1	0.900	4.9	0.900	6.7	0.700	8.5	0.500		
1.8	0.756	3.6	0.956	5.4	0.844	7.2	0.644	8.9	0.456		
[GRAF] CZAS W MIKROSEKUNDACH 21											
[TABELA] CZAS W MIKROSEKUNDACH 180											

ATRYBUT ALCOHOL, RED_WINE											
Val	**Sim**	**Val**	**Sim**	**Val**	**Sim**	**Val**	**Sim**	**Val**	**Sim**	**Val**	**Sim**
8.4	0.446	9.4	0.600	10.1	0.708	11	0.846	11.95	0.992	13	0.846
8.5	0.462	9.5	0.615	10.2	0.723	11.07	0.856	**12**	**1**	13.1	0.831
8.7	0.492	9.55	0.623	10.3	0.738	11.1	0.862	12.1	0.985	13.2	0.815
8.8	0.508	9.567	0.626	10.4	0.754	11.2	0.877	12.2	0.969	13.3	0.800
9	0.538	9.6	0.631	10.5	0.769	11.3	0.892	12.3	0.954	13.4	0.785
9.05	0.546	9.7	0.646	10.55	0.777	11.4	0.908	12.4	0.938	13.5	0.769
9.1	0.554	9.8	0.662	10.6	0.785	11.5	0.923	12.5	0.923	13.57	0.759
9.2	0.569	9.9	0.677	10.7	0.800	11.6	0.938	12.6	0.908	13.6	0.754
9.233	0.574	9.95	0.685	10.75	0.808	11.7	0.954	12.7	0.892	14	0.692
9.25	0.577	10	0.692	10.8	0.815	11.8	0.969	12.8	0.877	14.9	0.554
9.3	0.585	10.03	0.697	10.9	0.831	11.9	0.985	12.9	0.862		
[GRAF] CZAS W MIKROSEKUNDACH 22											
[TABELA] CZAS W MIKROSEKUNDACH 666											

ATRYBUT NUM_HREFS, ONLINE_NEWS_POPULARITY											
Val	**Sim**	**Val**	**Sim**	**Val**	**Sim**	**Val**	**Sim**	**Val**	**Sim**	**Val**	**Sim**
0	0.984	23	0.941	46	0.865	69	0.789	92	0.714	124	0.609
1	0.987	24	0.938	47	0.862	70	0.786	93	0.711	127	0.599
2	0.990	25	0.934	48	0.859	71	0.783	94	0.707	140	0.556
3	0.993	26	0.931	49	0.855	72	0.780	96	0.701	142	0.549
4	0.997	27	0.928	50	0.852	73	0.776	97	0.697	143	0.546
5	**1**	28	0.924	51	0.849	74	0.773	98	0.694	145	0.539
6	0.997	29	0.921	52	0.845	75	0.770	100	0.688	148	0.530
7	0.993	30	0.918	53	0.842	76	0.766	101	0.684	150	0.523
8	0.990	31	0.914	54	0.839	77	0.763	102	0.681	152	0.516
9	0.987	32	0.911	55	0.836	78	0.760	103	0.678	153	0.513
10	0.984	33	0.908	56	0.832	79	0.757	104	0.674	158	0.497
11	0.980	34	0.905	57	0.829	80	0.753	105	0.671	159	0.493
12	0.977	35	0.901	58	0.826	81	0.750	106	0.668	161	0.487
13	0.974	36	0.898	59	0.822	82	0.747	107	0.664	162	0.484
14	0.970	37	0.895	60	0.819	83	0.743	108	0.661	171	0.454
15	0.967	38	0.891	61	0.816	84	0.740	110	0.655	186	0.405
16	0.964	39	0.888	62	0.813	85	0.737	116	0.635	187	0.401
17	0.961	40	0.885	63	0.809	86	0.734	117	0.632	304	0.016
18	0.957	41	0.882	64	0.806	87	0.730	118	0.628		
19	0.954	42	0.878	65	0.803	88	0.727	119	0.625		
20	0.951	43	0.875	66	0.799	89	0.724	120	0.622		
21	0.947	44	0.872	67	0.796	90	0.720	122	0.615		
22	0.944	45	0.868	68	0.793	91	0.717	123	0.612		
[GRAF] CZAS W MIKROSEKUNDACH 38											
[TABELA] CZAS W MIKROSEKUNDACH 11227											

ATRYBUT NUM_IMGS, ONLINE_NEWS_POPULARITY											
Val	**Sim**	**Val**	**Sim**	**Val**	**Sim**	**Val**	**Sim**	**Val**	**Sim**	**Val**	**Sim**
0	0.977	16	0.898	32	0.773	48	0.648	64	0.523	90	0.320
1	0.984	17	0.891	33	0.766	49	0.641	65	0.516	91	0.313
2	0.992	18	0.883	34	0.758	50	0.633	66	0.508	92	0.305
3	**1**	19	0.875	35	0.750	51	0.625	67	0.500	93	0.297
4	0.992	20	0.867	36	0.742	52	0.617	68	0.492	98	0.258
5	0.984	21	0.859	37	0.734	53	0.609	69	0.484	99	0.250
6	0.977	22	0.852	38	0.727	54	0.602	70	0.477	100	0.242
7	0.969	23	0.844	39	0.719	55	0.594	71	0.469	101	0.234
8	0.961	24	0.836	40	0.711	56	0.586	73	0.453	108	0.180
9	0.953	25	0.828	41	0.703	57	0.578	75	0.438	111	0.156
10	0.945	26	0.820	42	0.695	58	0.570	76	0.430	128	0.023
11	0.938	27	0.813	43	0.688	59	0.563	77	0.422		
12	0.930	28	0.805	44	0.680	60	0.555	79	0.406		
13	0.922	29	0.797	45	0.672	61	0.547	80	0.398		
14	0.914	30	0.789	46	0.664	62	0.539	83	0.375		
15	0.906	31	0.781	47	0.656	63	0.531	84	0.367		
[GRAF] CZAS W MIKROSEKUNDACH 35											
[TABELA] CZAS W MIKROSEKUNDACH 12795											

Jakie informacje chcieliśmy uzyskać?

- W zbiorze *ForestFires*, informację o tym, jakie jest podobieństwo wszystkich rekordów reprezentujących pożary lasów w stosunku do rekordu reprezentującego pożar, który wystąpił przy temperaturze powietrza 25 stopni Celsjusza.
- W zbiorze *ForestFires*, informację o tym, jakie jest podobieństwo wszystkich rekordów reprezentujących pożary lasów w stosunku do rekordu reprezentującego pożar, który wystąpił przy wietrze o prędkości 4 km/h.
- W zbiorze *RedWine*, informację o tym, jakie jest podobieństwo wszystkich rekordów reprezentujących jakość czerwonych win w stosunku do rekordu reprezentującego wina o 12-procentowej zawartości alkoholu.
- W zbiorze *OnlineNewsPopularity*, informację o tym, jakie jest podobieństwo wszystkich rekordów reprezentujących informacje o publikowanych artykułach w stosunku do rekordu reprezentującego artykuł, w którym znalazło się 5 linków.
- W zbiorze *OnlineNewsPopularity*, informację o tym, jakie jest podobieństwo wszystkich rekordów reprezentujących informacje o publikowanych artykułach w stosunku do rekordu reprezentującego artykuł, w którym znalazły się 3 zdjęcia.

Powyżej przedstawiono wyniki dla operacji na różnej wielkości zbiorach. Dla oszczędności miejsca przedstawiono wartości podobieństwa dla niezduplikowanych wartości z konkretnych atrybutów.

Można zauważyć, że im większy zbiór, tym oszczędność czasowa jest coraz większa. Dla zbioru danych *ForestFires* liczącego 517 rekordów, obliczenie podobieństwa dla atrybutu z mniejszą ilością duplikatów jest **2.3 razy** korzystniejsze niż wyliczenie podobieństwa w tabeli. Z kolei dla zbioru z większą ilością duplikatów wynik jest już **8.5 razy** lepszy.

Dla zbioru *RedWine*, który liczy już 1599 rekordów, obliczenie podobieństwa dla atrybutu *Alkohol* jest wykonywane w grafie **30 razy** szybciej w stosunku do operacji wykonywanych w tabeli.

Jednakże prawdziwą korzyść widać na dużo większym zbiorze *OnlineNewsPopularity*, liczącym blisko 40 tysięcy rekordów. Podczas liczenia podobieństwa dla atrybutu reprezentującego liczbę linków w artykule, obliczenia w grafie są **290 razy** bardziej wydajniejsze, niż w tabeli. Dla atrybutu mówiącego, ile zdjęć znajduje się w artykule, otrzymujemy **365 razy** lepszy wynik.

Dla największego zbioru liczącego około 40 tysięcy rekordów, oszczędność czasowa jest gigantyczna. Wynika to z faktu, iż w zbiorze tym znajduje się mnóstwo duplikatów, co automatycznie pomniejsza ilość operacji potrzebnych do wyliczenia podobieństwa w grafie. **Im większy zbiór danych i im więcej duplikatów w poszczególnych atrybutach, tym oszczędność czasu przy obliczaniu podobieństwa jest większa!**

4. Podsumowanie

Przeprowadzone eksperymenty pokazują, jak ogromną korzyścią jest przechowywanie danych w asocjacyjnych strukturach danych AGDS. Większość informacji jest dostępna właściwie bez jakiegokolwiek nakładu czasowego. Bardzo dużo informacji jesteśmy w stanie uzyskać w czasie dużo bardziej korzystnym niż ten, który jest konieczny do wykonania obliczeń w tabeli – najpopularniejszej współcześnie strukturze do przechowywania danych.

Dzięki odpowiedniej budowie asocjacyjnego grafu, bardzo dużo informacji jest dostępnych dzięki samej jego implementacji. Mamy dostęp do posortowanych, niezduplikowanych list wartości dla każdego atrybutu, natychmiastowy dostęp do informacji o liście rekordów powiązanych z każdą konkretną wartością oraz w każdym rekordzie informacje o jego wartościach.

Informacje takie jak minimum, maksimum oraz zakres wartości w liście obiektów typu wartość, czy listy rekordów o określonej wartości dostępne są po wykonaniu operacji mających złożoność obliczeniową równą *O(1).*

Natomiast operacje bardziej skomplikowane, takie jak wyszukiwanie relacji koniunkcji i alternatywy, czy obliczanie podobieństwa w obrębie atrybutu wykonywane są dzięki algorytmom o złożoności obliczeniowej *O(n).*

Co najważniejsze, wszystkie przedstawione operacje są bardziej optymalne czasowo, jeżeli są wykonywane w asocjacyjnym grafie AGDS, niż te wykonywane w tabeli. W większości przypadków wyniki są tym lepsze, im więcej duplikatów znajduje się w liście wartości danego atrybutu.

Oczywiście graf umożliwia wyszukiwania wielu innych relacji, które są również wykonywane z dużą oszczędnością czasu, w porównaniu do operacji wykonywanych w tabelach. Relacje takie jak wyszukiwanie podobieństwa względem wszystkich atrybutów, czy wyszukiwanie obiektów podobnych do grupy obiektów, wykonywane jest przy pomocy większej złożoności obliczeniowej niż *O(n).* Nadal jednak wyszukiwanie takich relacji w grafie jest bardziej optymalne czasowo.

Podsumowując, w tym momencie wiele możliwości daje nam asocjacyjna struktura danych AGDS. Dzięki jej budowie jesteśmy w stanie wnioskować na temat różnego rodzaju relacji w bardzo korzystnym czasie. Wskazać jednak należy, iż jest ona bazą do dalszych rozważań w dziedzinie eksploracji danych. Obecnie istnieje możliwość przekształcenia struktury AGDS do postaci aktywnej sieci neuronowej AANG, gdzie jesteśmy w stanie uaktywniać poszczególne węzły w różnym czasie. Daje to możliwość pobudzania, hamowania i aktywowania innych węzłów (neuronów), w zależności od wag połączeń między nimi, co jest

kolejnym krokiem do jeszcze efektywniejszego i głębszego wnioskowania na temat relacji, uwzględniającego również różne zależności czasowe.

Bibliografia

[1] A. Horzyk. *Sztuczne systemy skojarzeniowe i asocjacyjna sztuczna inteligencja*. Akademicka Oficyna Wydawnicza EXIT, 2003

[2] S. Prata. *Język C++. Szkoła programowania. Wydanie VI.* Wydawnictwo Helion.

[3] *UCI Machine Learning Repository*. URL: *http://archive.ics.uci.edu/ ml/ datasets.html*.

[4] *Graphviz - Graph Visualization Software*. URL: *http://webgraphviz.com/*

[5] *draw.io v 7.2.9.* URL: *https://www.draw.io/*

[6] *Asocjacyjne grafowe struktury danych AGDS - Adrian Horzyk.* URL: *http://home.agh.edu.pl/~horzyk/lectures/miw/MIW-AGDS-%C4%86wiczenia.pdf* .

[7] *Struktury asocjacyjne oraz asocjacyjne grafy neuronowe do eksploracji wiedzy z danych – Adrian Horzyk.* URL: *http://home.agh.edu.pl/~horzyk/lectures/miw/MIW-AsocjacyjneGrafyNeuronowe.pdf* .

[8] Aho A. V.; Hopcroft A. E.; Ullman J. D. *Algorytmy i struktury danych*. Helion, 2003.

Spis rysunków

Spis tabel

Spis wzorów