W Y D A W N I C T W O P O L I T E C H N I K I Ś L Ą S K I E J W G L I W I C A C H

ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 2018
Seria: ORGANIZACJA I ZARZĄDZANIE z. 130

ANALIZA METOD PRZETWARZANIA INFORMACJI RUCHU SIECIOWEGO

Mateusz DMITRZAK[1], Kamil FIEDUKIEWICZ[2], Ireneusz J. JÓŹWIAK[3]

[1]Wydział Mechaniczny, Politechnika Wrocławska, Wrocław; dmitrzakmateusz@gmail.com
[2]Wydział Mechaniczny, Politechnika Wrocławska, Wrocław; kamil.fiedukiewicz@gmail.com
[3]Wydział Informatyki i Zarządzania, Politechnika Wrocławska, Wrocław; ireneusz.jozwiak@pwr.edu.pl

Streszczenie: W pracy przedstawiono weryfikację metod eksploracji danych stosowane do analizy ruchu sieciowego. Dokonano przeglądu oraz analizy porównawczej wykorzystywanych metod eksploracji danych w ruchu sieciowym. Przeprowadzono także analizę korzyści stosowania metod eksploracji danych w porównaniu ze standardowymi metodami analizy ruchu sieciowego.

Słowa kluczowe: ruch sieciowy, informacja, przetwarzanie, analiza.

DATA MINING METHODS ANALYSE OF NETWORK TRAFFIC INFORMATION

Abstract: The paper discusses data mining methods used to analyse the network traffic. A review and comparative analysis of data mining methods to analyse network traffic has been presented. Benefits of data mining methods were analysed and compared with standard methods of analysis of network traffic.

Keywords: network traffic, information, data mining, analyse.

1. Wprowadzanie

Analiza ruchu sieciowego jest zadaniem skomplikowanym ze względu na fakt bardzo dużej liczby danych, które muszą zostać przetworzone. Proces analizy danych pochodzących z logów ruchu sieciowego nie został dotychczas zautomatyzowany w sposób, w którym wiedza ekspercka do celów weryfikacji, poprawy uzyskiwanych rezultatów, nie byłaby wymagana.

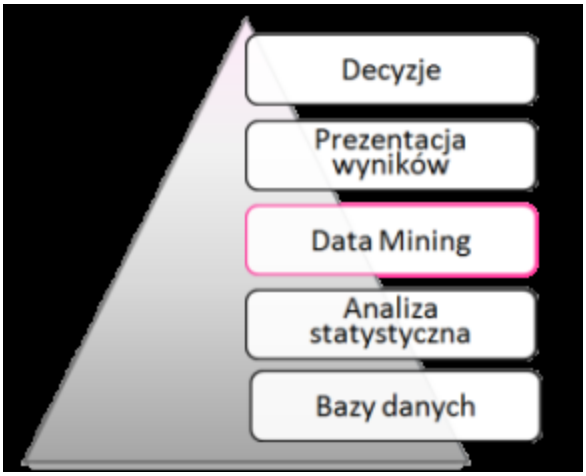
W artykule (Daszczuk et al., 2000) zostały przedstawione podstawowe metody analizy ruchu sieciowego wykorzystujące do tego celu wiedzę eksperta. Analiza ta opiera się na wykorzystaniu doświadczenia danej osoby, czy też systemu eksperckiego, do celów przetwarzania zebranych danych. Jest to metoda efektywna lecz mało wydajna i wymaga zaangażowania wielu ekspertów. Inne metody, częściowo zautomatyzowane, zostały zaprezentowane w pracach (Fayyad, 1996), (Miskov, 20.11.2018), (Silwattananusarn et al., 2016). Wykorzystują one dotychczas zebraną oraz usystematyzowaną wiedzę w postaci baz danych do celów analizy zebranych informacji. Podejście to opisane jest w (Fayyad, 1996) jako wysoce wydajne, lecz nie wskazano nowych relacji pomiędzy danymi.

Kolejnym podejściem jest wykorzystanie do celów analizy ruchu sieciowego metod eksploracji danych (ang. data mining) (Gawrysiak, and Okoniewski, 2000b), (Lee, 20.11.2018), czyli procesu selekcji, eksploracji i modelowania dużych ilości danych, które służą do badania nieznanych regularności i związków występujących w danych. Umożliwia to nie tylko automatyzację procesu analizy danych ruchu sieciowego, ale przede wszystkim umożliwia znajdowanie, początkowo nie znanych, użytecznych w procesie dalszego przetwarzania, relacji pomiędzy danymi. Celem artykułu jest dokonanie przeglądu oraz przeprowadzenie analizy metod przetwarzania informacji ruchu sieciowego wykorzystujących eksplorację danych do celów zbadania danych ruchu sieciowego.

2. Sztuczna inteligencja w eksploracji danych

W pracy (Fayyad, 1996) sztuczna inteligencja jest opisywana jako nauka obejmująca zagadnienia logiki rozmytej, obliczeń ewolucyjnych, sieci neuronowych, automatyki i robotyki. Sztuczna inteligencja to dział informatyki, którego przedmiotem jest badanie reguł rządzących inteligentnymi zachowaniami człowieka, tworzenie modeli formalnych tych zachowań i programów komputerowych symulujących te zachowania.

Dziedziną, wykorzystującą osiągnięcia sztucznej inteteligencji, a zajmującą się zagadnieniem analizowania danych, jest ekspoloracja danych. W artykule (Gawrysiak, and Okoniewski, 2000a) eksploracja danych została przedstawiona jako przeszukiwanie i analiza zbiorów danych w celu znalezienia związków między nimi w taki sposób, by były one zarówno zrozumiałe jak i przydatne.



Rysunek 1. Eksploracja danych (Gawrysiak, and Okoniewski, 2000a).

Na rysunku 1 został przedstawiony schemat etapów analizy danych z zaznaczonym miejscem procesu ich eksploracji.

3. Eksploracja danych

W pracy (Fayyad, 1996) Michael Fayyad i inni (1996) zdefiniowali eksplorację danych jako proces selekcji, przetwarzania i modelowania dużych liczby danych, który służy odkrywaniu regularności i związków występujących w nich, nieznanych początkowo. Celem tego procesu jest uzyskanie wyników użytecznych dla właściciela danych. Eksploracja danych jest to więc proces, w którego skład wchodzi kilka etapów. Proces ten pozwala na odkrywanie związków i regularności zachodzących między danymi, które nie są ani wyraźne, ani oczywiste. Nie tylko dlatego, że analizowane dane charakteryzują się zarówno znaczną objętością, jak i często również wielowymiarowością. Także dlatego, iż wykorzystanie i łączenie metod analiz pochodzących z różnych dziedzin, jak np. statystyka matematyczna i sztuczna inteligencja, pozwala na formułowanie problemów, z którymi każda z tych metod samodzielnie nie potrafiłaby udzielić satysfakcjonujących odpowiedzi. Odpowiedzi, jakie są uzyskane w wyniku eksploracji danych, należy traktować raczej jako przypuszczenia niż kategoryczne stwierdzenia. Jednakże przy odpowiednim sformułowaniu pytania, dają one cenne informacje.

W artykule (Daszczuk et al., 2000) autorzy zauważyli, że podstawowym sposobem uzyskiwania wiedzy z baz danych jest zadawanie zapytań, właściwe opracowywanie odpowiedzi i przedstawianie ich w formie raportów. W ten sposób można, dla przykładu poznać dane osobowe wszystkich klientów kupujących produkt A i B, co najmniej jeden raz.

Aby zadać pytanie, należy mieć świadomość o istnieniu dokładnie tego związku między tymi produktami, a więc posługiwać się pewną wiedzą a priori. Techniki eksploracji danych pozwalają zaś na odkrycie korelacji, których nie trzeba precyzyjnie definiować w momencie przeprowadzania analiz. Należy jedynie założyć występowanie tylko ogólnego rodzaju zależności, nie ograniczającego się do wskazywania konkretnych produktów. Czasem eksploracja danych kojarzona jest też ze statystyką matematyczną, a więc warto jeszcze wspomnieć o związku między tymi dwoma sposobami analizy danych. Poszczególne techniki eksploracji danych zawierają elementy analiz statystycznych. W publikacjach (Gawrysiak, and Okoniewski, 2000b), (Lee, 20.11.2018) eksploracja danych, jako proces, skupia się na analizie dużych ilości danych. Z powodu ograniczeń aplikacji, w wielu zastosowaniach nie jest możliwa analiza lub nawet dostęp do całej bazy danych. Dlatego też wymagane jest próbkowanie danych w celu wyboru danych reprezentatywnych. Musi ono być przeprowadzone w taki sposób, aby uwzględniać cele analiz. Często to jednak wyklucza stosowanie tradycyjnych metod statystycznych. W celu zapobiegania błędów występujących w modelu decyzyjnym, powstałych w wyniku analiz eksploracji danych, w przypadku, np. defraudacji, z bazy danych należy pobrać dużą liczbę obserwacji zawierających takie przypadki. Wynika to z samej natury zjawiska: defraudacja występuje dosyć rzadko w stosunku do przypadków prawidłowych, w związku z czym wybór grupy reprezentatywnej w sposób wyłącznie statystyczny może bardzo łatwo wyeliminować z analiz obserwacje zawierające defraudacje. Bazy danych mogą przechowywać dane w formie, które nie odpowiadają zastosowaniom do celów statystycznych. W takim przypadku poszukiwane są inne metody analiz, będące często także po prostu pewną modyfikacją metod statystycznych. Wyniki procesu eksploracji danych należy konfrontować z rzeczywistością i tylko dzięki powiązaniu z wiedzą ekspercką można ocenić ich prawidłowość. Statystyka zaś, z założenia stara się zdefiniować jeden model, który niezależnie od szczególnego przypadku analizowanych danych, daje pożądane rezultaty. Na koniec warto dodać, że dane do celów analiz statystycznych mogą pochodzić z badań eksperymentalnych, podczas gdy analizy eksploracji danych zawsze skupiają się na danych rzeczywistych.

4. Eksploracja danych w metodach analizy ruchu sieciowego

Dla dużych sieci komputerowych potrzeba monitorowania ruchu sieciowego stała się koniecznością. Zaproponowano nowe podejście do zbierania danych sieciowych oraz analizy ruchu sieciowego. W artykule (Daszczuk et al., 2000) przedstawiono główne problemy monitorowania sieci o dużej przepustowości. Dla przykładu łącze o przepustowości 10 Gb/s umożliwia przesłanie ponad 100 TB danych w ciągu dnia, co powoduje olbrzymie problemy w sposobie magazynowania i przetworzenia takiej ilości danych. Rozważa się dwa główne

problemy ze względu na szybkość analizy danych ruchu sieciowego. Pierwszym z nich jest infrastruktura, która byłaby w stanie zmagazynować tyle danych, musiałaby być bardzo zaawansowana oraz kosztowna. Z drugiej strony odkrywanie informacji ze zgromadzonych danych jest również wymagające obliczeniowo i czasochłonne. W celu zmniejszenia liczby gromadzonych, danych autorzy rozważają dwa następujące podejścia opisywane w (Daszczuk et al., 2000). Próbkowanie pakietów polega na tym, że rejestrowany jest tylko podzbiór ze wszystkich przepływających pakietów. Dla przykładu, jest to jeden z określonej liczby N pakietów (np. co dziesiąty). Analiza przepływów, to podejście polegające na przypisaniu pakietu do przepływu, który charakteryzuje się tymi samymi wartościami pól w pakiecie nagłówka pakietu, np. adres IP źródła, celu, czy numeru portów źródła i portu docelowego. Jednym z przykładów wdrożenia metody próbkowania pakietów, wykorzystującej do tego celu eksplorację danych jest technologia NetFlow przedstawiona w (Gawrysiak, and Okoniewski, 2000b), (Patterson, 02.09.2009). Technologia NetFlow została opracowana przez firmę Cisco Systems i jest podstawą gromadzenia kompletnych statystyk dotyczących ruchu sieciowego na interfejsach aktywnych urządzeń sieciowych (przełączniki, routery). Protokół NetFlow zaimplementowany jest w sieciowym systemie operacyjnym o nazwie IOS (ang. Internetwork Operating System) urządzeń sieciowych. Najczęściej wykorzystywany jest na interfejsach routerów brzegowych. Zebrane dane sumaryczne strumieni sesji host – host oraz wybranych interfejsów eksportowane są do tzw. NetFlow Collector'a. Poniżej omówiono zasadę działania i sposoby wykorzystywania danych zbieranych przy pomocy technologii NetFlow.

5. Analiza metod przetwarzania informacji

W pracy (Gawrysiak, and Okoniewski, 2000a) zostały wymienione technologie gromadzenia danych w oparciu o wymienione powyżej sposoby redukcji rejestrowanych danych z ruchu sieciowego. Ze względu na szybkość oraz efektywność analizy, mamy do czynienia z technologiami: NetFlow, RTFM oraz sFlow będącymi połączeniem próbkowania pakietów i identyfikacji przepływu. Jednocześnie zostało zaproponowane podejście rejestrowania danych z przepływu. Moduł NetLogger, będący elementem całego pakietu Analyzer, w którego skład wchodzi także moduł NetMiner gromadzi dane w postaci rekordów, które następnie poddawane są analizie. Analiza, z wykorzystaniem modułu NetMiner, w głównej mierze została oparta na algorytmach odkrywania reguł asocjacyjnych. Włoscy naukowcy, jako jeden z przykładów zastosowania eksploracji danych do analizy ruchu sieciowego, proponują możliwość odnajdywania aplikacji peer-to-peer w sieci (w skrócie P2P). Ze względu na fakt, że aplikacje są instalowane i kontrolowane bezpośrednio przez samych użytkowników sieci bardzo ważny dla administratorów jest nadzór nad ruchem tego typu. Propozycją rozwiązania

problemu identyfikacji ruchu P2P jest eksploracja danych w oparciu o reguły asocjacji wiążące ze sobą wykorzystywane przez te aplikacje porty.

Znaczącą rolę gromadzenia danych z ruchu sieciowego z wykorzystaniem NetFlow opisuje Michael Patterson (2009) w (Patterson, 02.09.2009). Autor przedstawia NetFlow jako "sieciową konieczność" dla przedsiębiorstw i możliwość wykorzystania technologii w celu analiz, planowania, monitorowania sieci, profilowania, hurtowni danych i eksploracji danych. Proponowane są rozwiązania NetQoS oraz Plixer służące do monitorowania danych. NetFlow z kolei określane jest jako jedna z najlepszych metod kontrolowania tego, co dzieje się w sieci. Zalety wykorzystania eksploracji danych w odniesieniu do danych zgromadzonych w oparciu o NetFlow prezentowane są również w (Daszczuk et al., 2000), (Świątek, 20.11.2018). Autorzy artykułu (Daszczuk et al., 2000) podkreślają, że problem analizy danych NetFlow jest niezmiernie trudny. Z drugiej strony argumentują, że przy wykorzystaniu odpowiednich narzędzi rezultaty mogą pomóc w identyfikacji spamu, ataków i w wykrywaniu zagrożeń. Jako główny obszar wykorzystania takiego podejścia wskazano wykorzystanie go przez dostawców Internetu, w celu dbałości o bezpieczeństwo sieci. W artykule zostały przedstawione przykłady wykorzystania analiz eksploracji danych, w celu wykrywania anomalii związanych z bezpieczeństwem sieci. Na przykład bardzo duża liczba połączeń z jednej maszyny (hosta) może wskazywać na atak typu DoS (ang. Denial of Service). Oznacza to atak polegający na uniemożliwieniu działania poprzez wykorzystanie zasobów. Inny atak ma nazwę „worm" (pol. robak) i jest to mały szkodliwy program samodzielnie rozprzestrzeniający się po sieci). Tematykę wykorzystania metod eksploracji danych sieciowych w obszarze wykrywania zagrożeń porusza Wenke Lee i in. (2018) (Lee, 20.11.2018). Opisuje powstawanie modeli wykrywania zagrożeń w oparciu o użycie technik eksploracji danych oraz zastosowanie zaproponowanych rozwiązań dla systemów Real-Time IDS (ang. Real-Time System – system czasu rzeczywistego) oraz systemu IDS (ang. Intrusion Detection System – system wykrywania zagrożeń).

Jeżeli celem badań jest odkrywanie wpływu zachowań użytkowników i ich preferencji na ruch sieciowy, to mamy do czynienia z problemem zbliżonym do klasycznych zastosowań eksploracji danych, takich jak segmentacja klientów, czy badanie „koszyka zakupów", przedstawionych w (Miskov, 20.11.2018). Do takich analiz wykorzystuje się dane opisujące cechy użytkowników wraz z danymi o działaniu sieci. Można w ten sposób szukać reguł, korelacji i prawidłowości dotyczących na przykład godzin pracy, pracy w określonych podsieciach, czy liczby przesyłanych pakietów przez określone grupy użytkowników. W efekcie otrzymuje się, np. reguły asocjacyjne, które oznaczają, że kierownicy wyższego szczebla logują się do sieci jedynie z własnego terminala, podobnie jak grupy użytkowników otrzymane przez klasteryzację lub jako liście drzewa decyzyjnego, lub jako zbiory najważniejszych atrybutów użytkownika, w metodologii zbiorów przybliżonych. Wiedza taka może posłużyć z pewnością administratorom sieci do efektywniejszego zarządzania użytkownikami i zbiorami ich uprawnień.

Jeżeli przedmiotem analizy jest działanie infrastruktury technicznej sieci, to mamy do czynienia z problemem, do którego rozwiązania wykorzystywane są statystyki sieciowe. Dzięki statystykom sieciowym odkrywane są prawidłowości w działaniu elementów danej sieci. Można w ten sposób szukać wszelkich anomalii działania systemu, takich jak „wąskie gardła" powodujące zwolnienie lub zablokowanie transmisji oraz próbować przewidywać awarie elementów sieci, które charakteryzowane są przez poprzedzające je objawy. Przy takich zastosowaniach występują jednak istotne ograniczenia. Przede wszystkim, aby przewidzieć anomalię działania sieci lub jej awarię, musimy posiadać wystarczająco dużo danych. Awarie nie zdarzają się często, więc aby otrzymać dane o liczbie awarii wystarczających do zbudowania modelu predykcyjnego, potrzeba często odfiltrować logi operacji sieciowych. Kolejnym problemem jest tu stosowanie kryterium nowości wiedzy odkrytej metodami KDD (ang. Knowledge Discovery in Databases) (Daszczuk et al., 2000), (Świątek, 20.11.2018). W ostatnim etapie cyklu KDD uzyskana wiedza jest oceniana przez ekspertów danej dziedziny, a wnioski przekazywane na zasadzie sprzężenia zwrotnego na pierwszy etap kolejnego cyklu. W danych automatycznie wygenerowanych przez elementy sieci, zdobyte wyniki są w większości oceniane przez ekspertów jako trywialne lub uprzednio znane. Często znajdują oni jednak interesujące wyniki lub wnioski na potwierdzenie swoich intuicji nabytych przez doświadczenie. Eksperci uznają, według badań naszego zespołu, reguły dotyczące działania sieci za interesujące kilka razy rzadziej niż np. w wiedzy dotyczącej użytkowników. Ma to z pewnością źródło w charakterze danych, które dla opisu elementów sieciowych podlegają znacznie bardziej ścisłym regułom i ograniczeniom. Wśród danych generowanych przez grupy ludzi – użytkowników, KDD odkrywa często nowe prawidłowości natury psychologicznej i socjologicznej. Prawa rządzące aparaturą sieciową są zazwyczaj dokładnie opisane w dokumentacji technicznej i znane ekspertom. Dlatego właśnie odkrywanie wiedzy w danych automatycznie wygenerowanych wymaga większej liczby danych wejściowych i większej selektywności metod. Mimo ściśle określonych zasad działania istnieją wciąż reguły, które nie są uwzględnione w specyfikacjach, mogące wynikać, np. z efektów współdziałania różnych elementów sieci. Takie nowe reguły i zależności często pozwalają na lepszą diagnostykę i optymalizację działania sieci lub w celu lepszego zaprojektowania jej nowych elementów.

Każda nowoczesna firma telekomunikacyjna posiada wiele zasobów sieciowych zarówno w postaci sieci telekomunikacyjnej jak i komputerowej. Korzystają z nich miliony abonentów. Poprawne działanie tych sieci jest warunkiem niezbędnym jej dla funkcjonowania i zyskowności.

Pierwszy eksperyment, dotyczący wyszukiwania nieprawidłowości działania sieci na podstawie logów routerów w sieci korporacyjnej, był czystym przypadkiem analizy danych generowanych automatycznie. Parametry dostępne w logach routerów Cisco (Cisco Systems Inc., 20.11.2018) były analizowane głównie za pomocą reguł asocjacyjnych. Awarie, które były celem badania, zdarzają się jednak tak rzadko, że stanowiły one w otrzymanych danych

mało znaczący odsetek. Dlatego prawie wszystkie otrzymane reguły zostały przez ekspertów zaklasyfikowane jako „znane" lub „trywialne". Był to pierwszy w serii eksperymentów wykonywanych przez zespół, a oprócz merytorycznego niepowodzenia przyniósł pierwsze efekty dotyczące porównań narzędzi KDD.

Znacznie lepsze efekty przyniósł eksperyment mający na celu przewidywanie wielkości ruchu w sieci komórkowej. Można go zakwalifikować do kategorii wyszukiwania wiedzy z danych generowanych przez użytkowników systemów sieciowych. Zbudowano model predykcyjny, który następnie był udoskonalany poprzez dodawanie nowych atrybutów oraz stosowanie clusteringu jako ostatniego etapu obróbki wstępnej danych. Pierwotnie, model oparty jedynie o regresję wieloraką bazującą na danych o rodzajach terenu z bazy typu GIS był obarczony dużym, nieakceptowalnym dla ekspertów błędem. Poprzez kilkukrotne zbudowanie modelu zgodnie z zasadami cyklu KDD po uwzględnieniu za każdym razem opinii ekspertów, udało się zmniejszyć błąd trzykrotnie, co pozwoliło zastosować model w planowaniu sieci komórkowej.

Kolejne dwa eksperymenty opisane w (Daszczuk et al., 2000), (Świątek, 20.11.2018), to przewidywanie nieprawidłowości w działaniu sieci komórkowej oraz analiza sąsiedztw elementów w sieci komórkowej. Oba one wykonywane były na podobnych próbach danych zawierających informacje o działaniu elementów sieci komórkowej. Jest to więc przypadek wyszukiwania wiedzy w danych wygenerowanych automatycznie. Przy zastosowaniu podobnych metod, takich jak reguły asocjacyjne, drzewa decyzyjne, metody wizualizacyjne uzyskiwane reguły w większości potwierdzały co najwyżej wiedzę i doświadczenie ekspertów. Jednakże lepsze wyniki odnotowane zostały w przypadku eksperymentu uwzględniającego sąsiedztwa elementów sieciowych. Dane te nie są generowane automatycznie, są zaś wynikiem położenia geograficznego stacji bazowych oraz topologii sieci. Poprawa wyników w tym przypadku potwierdziła tezę o mniejszej przydatności danych automatycznie wygenerowanych dla odkrywania wiedzy metodami KDD.

6. Podsumowanie

Obecnie można zauważyć znaczny rozwój metod i narzędzi do automatycznego analizowania zasobów sieciowych różnego typu. Trudno jednak na razie przewidzieć czy rozwój tych technologii pójdzie w kierunku wykorzystania obecnych metod dla odpowiednio spreparowanych danych z sieci, czy też powstaną nowe metody przystosowane jedynie do analizy sieci.

Rozwój metod analizy ruchu sieciowego spowodowany jest faktem, że w czasach powszechnego dostępu do systemów sieciowych, przewagę osiągnie ten, kto potrafił będzie z niej pobierać nie tylko dane, ale i to co najcenniejsze: wiedzę.

Wykorzystanie eksploracji danych w celu analizy ruchu sieciowego może się przyczynić do poprawy efektywności analizowania informacji. W artykule zostały przedstawione podstawowe metody sztucznej inteligencji oparte o eksplorację danych (ang. data minig), które umożliwiają wykorzystanie metod „inteligentnych" do zastosowania w sieciach komputerowych. Wpływ omawianych metod w stosunku do standardowych analiz powoduje znacznie przyspieszenie przetwarzania zebranych danych. Wykorzystanie eksploracji danych do celów analizy ruchu sieciowego poprawia obraz zebranych danych, co wpływa na poprawę możliwości ich analizy.

Bibliografia

1. Cisco Systems Inc. (20.11.2018). Available online: https://www.cisco.com/c/pl_pl /products/index.html
2. Daszczuk, W., Gawrysiak, P., Gerszberg, T., Kryszkiewicz, M., Mieścicki, J., Muraszkiewicz, M., Okoniewski, M., Rybiński, H., Traczyk, T., Walczak, Z. (2000). *Data mining for Technical Operation of Telecommunications Companies*: *a Case Study*. Paper presented at 4th World Multiconference on Systemics, Cybernetics and Informatics SCI'2000, Orlando, Florida.
3. Fayyad, U.M. Piatetsky-Shapiro, G., Smyth, P. (1996). *From Data Mining to Knowledge Discovery in Databases*. Communications of the ACM 1996, 39, 37-54. doi: 10.1145/230798.358186
4. Gawrysiak, P., Okoniewski, M. (2000). Applying Data Mining Methods for Cellular Radio Network Planning. In M. Kłopotek, M. Michalewicz, S.T. Wierzchoń, *Proceedings of the IIS'2000 Symposium*, Bystra, Poland, June 12-16, 2000 (pp. 87-98). Warsaw: Physica-Verlag Heidelberg
5. Gawrysiak, P., Okoniewski, M., (2000). *Knowledge Discovery in the Internet*. Archiwum Informatyki Teoretycznej i Stosowanej, T. 12, z. 3, 203-233.
6. Lee, W., Stolfo, S.J. Mok, K.W. (20.11.2018). Adaptive Intrusion Detection: a Data Mining Approach. Retrived from wenke.gtisc.gatech.edu/papers/ai_review.ps
7. Miskov, N. (20.11.2018). Survey of tools for data mining. Retrived from home.etf.rs/~vm/cd1/papers/230.pdf
8. Patterson, M. (02.09.2009). NetFlow Analysis on the Cisco ASA. Retrived from https://www.lovemytool.com/blog/2009/09/plixer.html
9. Silwattananusarn, T., Kanarkard, W., Tuamsuk, K. (2016). Enhanced Classification Accuracy for Cardiotocogram Data with Ensemble Feature Selection and Classifier Ensemble. *Journal of Computer and Communications*, 4, 4.
10. Świątek, P. (20.11.2018). Available online www.math.uni.wroc.pl/~swiatek /word2art.doc