STUDIA INFORMATICA 2003
Volume 24 Number 2A (53)

Leszek BORZEMSKI
Politechnika Wrocławska, Instytut Sterowania i Techniki Systemów

DATA MINING DLA INTERNETU

Streszczenie. Analiza danych z wykorzystaniem metod i narzędzi data miningowych jest wykorzystywana w wielu zastosowaniach. Artykuł omawia możliwości wykorzystania data miningu na potrzeby analizy funkcjonowania podsystemu komunikacyjnego Internetu.

Słowa kluczowe: data mining, web mining, eksploracja danych, odkrywanie wiedzy, Internet.

DATA MINING MEETS INTERNET

Summary. Data mining is used in various application areas and now Internet research meets data mining. Data analysis based on statistical characteristics of collected data has exhausted its capabilities. It is proposed to apply different data mining methods and algorithms to gain our knowledge about the Internet.

Keywords: data mining, web mining, knowledge discovery, Internet.

1. Wprowadzenie

Data mining[1] (DM) jest podstawowym działaniem w procesie *odkrywania wiedzy w bazach danych* KDD (*Knowledge Discovery in Databases*) [11]. Jest to proces automatycznego odkrywania znaczącej, dotychczas nieznanej i pożytecznej wiedzy zawartej w dużych bazach danych, wiedzy ujawniającej ukryte własności badanej dziedziny. Poszukiwane są związki i ogólne prawidłowości, które są bezpośrednio niewidoczne i stąd

[1] W pracy, nawiązując do terminów Web mining i Internet mining, będziemy stosować termin data mining. Termin data mining tłumaczony jest w języku polskim różnie. Wśród wielu polskich terminów będących w użyciu, dwa sformułowania są popularne, a mianowicie *eksploracja danych* [14] i *drążenie danych* [12].

trudne do wykrycia. Wiedza odkrywana jest z wykorzystaniem zaawansowanych technik polegających na np. klasyfikowaniu, grupowaniu czy kojarzeniu. Uzyskaną wiedzę można wykorzystać do lepszego zrozumienia badanej rzeczywistości, ale także do podjęcia konkretnych decyzji biznesowych i technicznych [10, 11].

DM obejmuje coraz więcej obszarów zastosowań, przy czym najwięcej konkretnych rezultatów dotyczy procesów ekonomicznych i rynkowych. Rezultaty analiz DM pozwalają uzyskać m.in. istotne informacje o klientach i rynku, przez co umożliwiają dostosowanie działalności firmy do zmieniających się potrzeb i zwiększenie konkurencyjności. W efekcie pozwala to na wzrost zysku z prowadzonej działalności. Analiza koszyka i segmentacja klientów pozwalają na ukierunkowany marketing. Wykrywanie przestępstw i oszustw finansowych są częstymi zastosowaniami analizy danych prowadzonej metodami data miningowymi. Metody te są stosowane w badaniach naukowych, medycynie, farmakologii czy nawet sporcie. Przykładowo mają one zastosowanie do analizy danych klimatycznych, przewidywania pogody i klęsk żywiołowych (trzęsienia ziemi, huragany, powodzie czy wielkie pożary). DM stosuje się do analizy danych epidemiologicznych, w badaniach nad rakiem, do odkrywania nowych związków chemicznych i lekarstw, wykrywania nowych galaktyk i śladów życia w kosmosie. Obszarami zastosowań są telekomunikacja oraz systemy i sieci komputerowe, gdzie wykorzystuje się DM w celu wykrycia przestępstw telekomunikacyjnych i zagrożeń komputerowych, takich jak spamy, hakerzy, ataki DoS, a także do wykrywania obecności intruzów. DM znajduje również zastosowanie w diagnostyce i ocenie wydajności systemów technicznych [20]. Należy wspomnieć, że metody eksploracji danych tekstowych, tzw. *text mining*, są wykorzystywane przez wyszukiwarkę internetową Google. Text mining jest również wykorzystywany przez opracowaną w zespole autora wyszukiwarkę SearchSystem [3].

Generalnie, technik data miningowych możemy z powodzeniem użyć wszędzie tam, gdzie istnieją bardzo duże zbiory danych opisujących wybrany fragment rzeczywistości. Data mining możemy stosować alternatywnie do klasycznych technik analizy danych bazujących na metodach statystycznych oraz w takich sytuacjach, w których stosowanie tych metod analizy danych jest nieefektywne, niemożliwe, bądź nieuzasadnione. DM wykorzystuje ponadto możliwości twórczego działania, jakie daje udział człowieka w interakcyjnym przetwarzaniu i analizie danych. Do efektywnego stosowania DM niezbędne jest specjalistyczne oprogramowanie realizujące różne algorytmy DM i organizujące współpracę ze środowiskiem operacyjnym i użytkownikiem - specjalistą z danej dziedziny.

2. Internet mining v. Web mining

Internet rozwija się jako olbrzymia i rozproszona składnica zasobów informacyjnych i programowych przechowywanych w serwerach FTP oraz przede wszystkim na serwerach webowych. Internet widzimy dzisiaj jako źródło informacji na niemal wszystkie poszukiwane tematy. W celu uzyskania informacji korzystamy głównie z usługi WWW i odpowiednich systemów wyszukiwarek internetowych, np. takich jak Google czy AltaVista. Jakość funkcjonowania usługi WWW jest ściśle uzależniona od jakości pracy podsieci komunikacyjnej Internetu. Stąd wynika potrzeba ciągłego monitorowania pracy Internetu w zakresie jego funkcji komunikacyjnych i prowadzenia odpowiednich analiz z wykorzystaniem różnych podejść metodologicznych. Od szeregu lat wiele ośrodków prowadzi pomiary wydajności określonych połączeń internetowych, m.in. w celu dostarczenia informacji o aktualnych i predykowanych wartościach określonych wielkości charakteryzujących „stan zdrowia" (lub „pogodę") fragmentów Internetu (a właściwie jego systemu komunikacyjnego) [4, 6, 7, 13, 15, 17, 21]. Dane pomiarowe w tych badaniach są przetwarzane z wykorzystaniem klasycznych metod statystycznej analizy danych. Wprowadzenie metod data miningu daje szansę na uzyskanie nowych obiecujących rezultatów w zakresie analizy danych pomiarowych dotyczących Internetu.

Niniejsza praca wskazuje na wybrane możliwości wykorzystania data miningu do analizy danych charakteryzujących Internet w zakresie funkcjonowania jego podsystemu komunikacyjnego, a w szczególności jego struktury oraz wydajności i niezawodności. Tak rozumiane zastosowanie data mining dla Internetu określać będziemy dalej terminem *Internet mining* (IM)[1]. Internet mining będziemy odróżniać od zastosowań DM w badaniach sieci WWW, które związane są z terminem *Web mining*. Rozróżnienie to wynika z innych zakresów problemowych obu działań odkrywania wiedzy oraz dostępu do odmiennych baz danych i narzędzi pomiarowych. Należy jednak pamiętać o wzajemnym przenikaniu się obu zagadnień. Zastosowania metod DM dla systemów webowych dotyczą głównie e-biznesu, wyszukiwania informacji w sieci WWW oraz zarządzania siecią WWW (ośrodkami WWW) [18]. Analizie podlegają dzienniki przechowujące informacje o dostępach do serwerów webowych, a poszukiwane mogą być np. związki charakteryzujące użytkowników, ich zachowania oraz preferencje. Stosując techniki DM możemy próbować określić strukturę sieci WWW, klasy użytkowników odwiedzających określone serwisy WWW, czy też wyznaczać serwisy preferowane przez użytkowników. Na podstawie analizy DM

[1] Za takim terminem przemawia pochodzenie terminu Internet, który w sensie technicznym określa połączone sieci oparte na protokołach TCP/IP.

obserwowanego ruchu pakietów protokołu http możemy rozstrzygać np. o rozmieszczeniu serwerów webowych czy też lokalizacji zawartości stron WWW.

Web mining rozwija się znacznie szybciej aniżeli Internet mining. Jedną z przyczyn takiego stanu rzeczy jest to, że na potrzeby Web miningu dysponujemy w zasadzie nieograniczonymi zbiorami baz danych, z których naturalną bazę danych tworzy sama sieć stron WWW, którą na wiele sposobów eksplorują wszelakie wyszukiwarki i szperacze sieciowe. Potrzebne bazy danych do Web miningu możemy uzyskać w dość prosty sposób w przypadku serwera webowego bądź konkretnego ośrodka webowego. Bazy danych do Web miningu mogą być tworzone z dzienników serwerów webowych oraz zbiorów zawierających zapisy monitorowanego ruchu pakietów protokołu http. Bazy te w większości przypadków powstają w wyniku biernych eksperymentów pomiarowych w trakcie, których obserwujemy zachodzące zdarzenia bez żadnej celowej ingerencji w zachodzące procesy. Bazę taką możemy stworzyć również w trakcie aktywnego eksperymentu pomiarowego. Wówczas w wyniku pewnego celowego działania polegającego na generowaniu określonego ruchu do serwisu WWW, uzyskujemy odpowiedź systemu. W eksperymencie aktywnym możemy ukierunkować pomiary pod kątem potrzeb badanego problemu. Przykładem systemu, który realizuje aktywny eksperyment pomiarowy w celu oceny wydajności Weba po stronie klienta jest system Wing [1]. Bazy dotyczące Weba nie poddają się łatwo analizie klasycznymi metodami statystycznymi, co powoduje skierowanie zainteresowania ku metodom DM. Duża popularność Web miningu wynika także z tego, że rezultaty przeprowadzonej analizy bardzo często szybko można wykorzystać bezpośrednio w obszarze gospodarczym i przełożyć na odpowiednie działania marketingowe.

Na podstawie literatury dotyczącej metod i narzędzi analizy pomiarów podsieci komunikacyjnej sieci Internet można stwierdzić, że Internet mining, rozumiany w taki sposób, jak to przedstawiliśmy wcześniej, jest obszarem zastosowań data miningu, w którym odpowiednie prace pojawiają się dopiero w ostatnim okresie [2, 8, 9, 16].

3. Internet mining w Instytucie Sterowania i Techniki Systemów

Prace w zakresie Internet miningu rozwijane są w Zakładzie Rozproszonych Systemów Komputerowych instytutu. Ich celem jest badanie wybranych fragmentów Internetu pod kątem wydajności i funkcjonalności, a także odkrywanie ich struktury oraz detekcji występujących anomalii. Analiza danych pomiarowych prowadzona jest dla baz danych z pomiarami zgromadzonymi w trakcie zaplanowanych eksperymentów aktywnych. Pomiary realizowane są z wykorzystaniem przede wszystkim serwerów IBM RISC/6000 pracujących pod kontrolą systemów operacyjnych AIX. Wykorzystywane są także stacje MS Windows

2000 Server na komputerach PC. Opracowane oprogramowanie pomiarowe wykorzystuje własne procedury pomiarowe oraz korzysta z pewnych usług dostępnych w sieci. Analizy DM przeprowadzane są z wykorzystaniem systemu firmy IBM o nazwie Intelligent Miner for Data (IMD) [5, 19], który współpracuje z systemem bazy danych IBM DB2 i funkcjonuje na obu platformach operacyjnych. System ten w ocenie wielu zespołów badawczych jest jednym z najlepszych komercyjnych rozwiązań tego rodzaju. Jego pozytywną cechą jest bogactwo funkcji data mining i różnych algorytmów je realizujących, dostęp do szerokiego spektrum klasycznych statystycznych metod analizy danych, bogate środki do wizualizacji rezultatów analiz, a ponadto dostęp do interfejsu programowego API, dzięki czemu można w prosty sposób rozszerzać funkcjonalność systemu oraz budować własne systemy data mining. Aktualnie dostępna jest wersja 8.1, która zawiera polskojęzycznego klienta.

Podstawowym problemem było opracowanie metodologii prowadzenia aktywnych eksperymentów pomiarowych w Internecie oraz wybór metod i technik DM ze zbioru funkcji i technik oferowanych przez system IMD, które byłyby przydatne do rozwiązania sformułowanych przez nas problemów badawczych. Postawiono następujące wymagania wobec procesu gromadzonych danych: możliwość uzyskania wystarczająco dużej ilości danych, możliwość ustawienia niewielkich odstępów czasu pomiędzy pomiarami, możliwość monitorowania wielu parametrów pracy sieci, umożliwienie gromadzenia danych pochodzących z wielu źródeł oraz możliwość łatwego ich przekształcania (import i eksport). Wymagania wobec narzędzi pomiarowych dotyczyły: prostoty i szybkości działania, wytwarzania niewielkiego ruchu w sieci, niepowodowania zbytniego obciążenia ruterów i serwerów oraz umożliwienia bezproblemowego zapisu i gromadzenia rezultatów. Brano pod uwagę fakt, że pomiary aktywne są wykonywane przez wprowadzenie testowego, sztucznego ruchu do sieci i obserwację reakcji systemu komunikacyjnego. Pomiary tego rodzaju powodują dodatkowy ruch w sieci i mogą zniekształcić jej zachowanie, a przez to i rezultaty pomiarów. Kolejny problem, który występuje przy pomiarach wydajności sieci, polega na określeniu, z których bądź pomiędzy którymi węzłami sieci należy prowadzić pomiary. Węzły muszą być tak dobrane, aby można było przeanalizować badany problem. Wybór węzłów musi uwzględniać dostępność węzłów, możliwość oferowania przez węzeł określonej usługi wykorzystywanej przy pomiarach oraz jakość pracy węzła. Rozważane problemy były problemami definiowanymi pod kątem potrzeb użytkownika końcowego, stąd jednym z węzłów był węzeł z naszej lokalnej sieci komputerowej. Należało, oczywiście, określić miary oceny sieci. Do najważniejszych miar, które zostały uwzględnione w badaniach, należą: opóźnienie pakietów, przepustowość, dostępność, liczba utraconych pakietów, częstość otrzymywania wadliwych odpowiedzi oraz długość trasy przesyłu pakietu.

Faza zbierania danych musiała trwać dostatecznie długo, aby można było zbudować odpowiednio dużą bazę danych.

Tabela 1

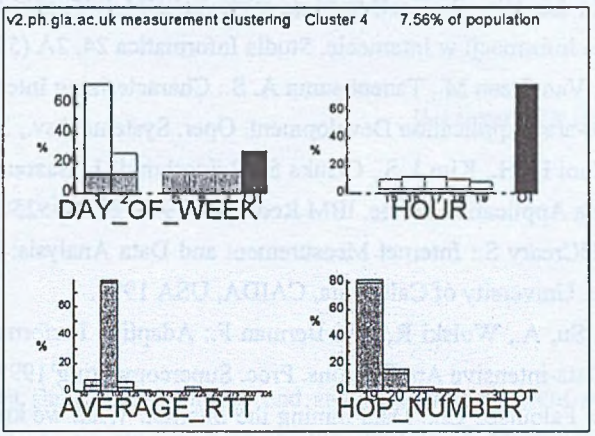
Wyniki badań

Lp.	Problem	Funkcje	Techniki	Rezultat
1.	Analiza zmienności czasu obiegu pakietu i liczby przebytych węzłów w zależności od chwili pomiaru	klasteryzacja klasyfikacja	sieci Kohonena drzewa decyzyjne	+
2.	Analiza zmienności rutingu	sekwencje wzorców		+
3.	Przewidywania czasu obiegu	predykcja klasyfikacja	sieci radialne drzewa decyzyjne	-
4.	Wykrywanie podobnych podsekwencji w przebiegach czasowych wartości średniego czasu obiegu	podobne sekwencje		+/-
5.	Wyszukiwanie sieci występujących na trasie przesyłu	asocjacje		+
6.	Określenie zależności między liczbą hopów a czasem obiegu	klasyfikacja	drzewa decyzyjne	+

W pracach wykorzystano m.in. następujące funkcje data miningu systemu IMD: klasteryzację, klasyfikację, asocjacje, predykcję, wzorce sekwencyjne, podobne sekwencje, w których zastosowano następujące techniki analizy, a mianowicie: drzewa decyzyjne, sieci neuronowe, algorytmy demograficzne oraz funkcje statystyczne. Wykaz problemów, które przeanalizowano z wykorzystaniem funkcji i technik data miningu dostępnych w systemie IMD przedstawia Tabela 1. Oznaczenie „+" w kolumnie „Rezultat" wskazuje na pozytywny wynik zastosowania DM w rozwiązywanym problemie, co oznacza, że zastosowana funkcja i technika analizy DM są skuteczne w odkrywaniu prawidłowości we wskazanym badaniu. Symbol "-„ oznacza, że badania przeprowadzone z wykorzystaniem określonej funkcji i techniki DM nie pozwoliły wyróżnić poszukiwanych reguł w konkretnym zbiorze danych analizowanym w określonym zadaniu badawczym. Natomiast symbol „+/-„ oznacza sytuację, że pomimo iż system generalnie odkrywał określone podobieństwa, to jednak w kilku przypadkach ich nie odkrył, mimo wyraźnych oznak ich występowania.

Przedstawimy teraz wykorzystanie funkcji klasteryzacji w problemie badania zmienności czasu obiegu pakietu i liczby przebytych węzłów w zależności od chwili pomiaru [2]. Klasteryzację z zastosowaniem sieci Kohonena przeprowadzono dla zbioru rekordów pomiarowych zawierających wyniki pomiarów czasów obiegu pakietu pomiędzy hostem zlokalizowanym w naszej sieci a wybranym hostem w Internecie. Pomiary były prowadzone co 30 minut przez okres trzech miesięcy. Po klasteryzacji otrzymano 16 klastrów. Najmniejszy z nich zawierał ok. 3% rekordów - największy 7,6%, przy czym większość

klastrów obejmowała po około 7% rekordów. W analizie brały udział następujące atrybuty: dzień tygodnia (day_of_week) i godzina (hour) wykonania pomiaru, średni czas obiegu pakietu (average_RTT) oraz liczba hopów (hop_number), która wyraża długość ścieżki od hosta wysyłającego pakiet do hosta docelowego. Charakterystykę drugiego co do wielkości klastra o numerze 4 zaprezentowano na rys. 1. Wykresy słupkowe przedstawiają częstości występowania określonych wartości atrybutu. Kontury narysowane linią przerywaną (w oryginale ciągłą linią o kolorze czerwonym) opisują rozkład wartości atrybutu w danym klastrze, podczas gdy szare słupki w tle pokazują rozkład wartości atrybutu w całym zbiorze danych. Rekordy przydzielone do tego klastra dotyczą pomiarów, które zostały wykonane prawie zawsze w poniedziałek (67% przypadków) lub we wtorek (24% przypadków). Reszta dni tygodnia wystąpiła sporadycznie. Rekordy z tego klastra odnoszą się przede wszystkim do pomiarów przeprowadzonych pomiędzy godziną 15, a 19, przy czym o każdej godzinie z tego zakresu dokonano nieco powyżej 10% wszystkich pomiarów należących do tej grupy. Na pozostałe dziewiętnaście godzin rozkłada się prawie 42% rekordów. Klaster charakteryzuje się tym, że w znacznej większości pomiarów (79%) średni czas obiegu należy do zakresu 75-100 ms, a długość trasy wynosi przeważnie 19 hopów (ponad 80% przypadków).



Rys. 1. Charakterystyka klastra o numerze 4
Fig. 1. Cluster 4

4. Uwagi końcowe

Przeprowadzone badania dotyczyły: zmienności czasu obiegu pakietu i liczby przebytych węzłów w zależności od chwili pomiaru, zmienności routingu, możliwości przewidywania czasu obiegu, występowania podobnych podsekwencji w przebiegach czasowych wartości średniego czasu obiegu, sieci występujących na trasie przesyłu oraz zależności pomiędzy

liczbą hopów a czasem obiegu. Należy podkreślić, że zasadniczym warunkiem powodzenia takich badań jest udział specjalistów od sieci komputerowych z dobrym zrozumieniem celów stosowania określonych metod DM, dzięki czemu można było prawidłowo przeprowadzić cały eksperyment, począwszy od zaplanowania pomiarów, a skończywszy na analizie wyników. Prace pokazały przydatność metod data miningu do analizy danych pomiarowych dotyczących wydajności oraz funkcjonalności fragmentów Internetu. Systemy data miningu mogą znaleźć zastosowanie w odkrywaniu pewnych zależności, których znajomość umożliwi polepszenie funkcjonowania oferowanych przez Internet usług komunikacyjnych.

LITERATURA

1. Borzemski L., Cichocki Ł., Nowak Z.: Wing – system do pomiaru wydajności usługi WWW po stronie klienta, Studia Informatica 24, 2A (53), 2003.
2. Borzemski L., Lubczyński Ł., Nowak Z.: Application of data mining for the analysis of Internet end-to-end performance, przekazane do opublikowania.
3. Borzemski L., Łopatka P.: Zastosowanie systemu IBM Inteligent Miner for Text do wyszukiwania informacji w Internecie. Studia Informatica 24, 2A (53), 2003.
4. Ballintijn G., Van Steen M., Tanenbaumn A. S.: Characterizing Internet Performance to Support Wide-area Application Development. Oper. Systems Rev., 34 (4), Oct. 2000.
5. Cabena P., Choi H. H., Kim I. S., Otsuka S., Reinschmidt J., Saarenvirta G.: Intelligent Miner for Data Application Guide. IBM Redbooks 1999, SG24-5252-00.
6. Claffy K., McCreary S.: Internet Measurement and Data Analysis: Passive and Active Measurement. University of California, CAIDA, USA 1999.
7. Faerman M., Su, A., Wolski R., and Berman F.: Adaptive Performance Prediction for Distributed Data-intensive Applications. Proc. Supercomputing'1999, 1999.
8. Faloutsos M., Faloutsos Ch.: Data-mining the Internet: What we know, what we don't, and how we can learn more. SIGCOMM'2002, 2002.
9. Garofalakis M., Rastogi R.: Data mining meets network management: The NEMESIS Project. Proc. of DMKD'2001, 2001.
10. Grossman R. L., Kamath Ch., Kegelmeyer P., Kumar V., Namburu R. R.: Data Mining for Scientific and Engineering Applications. Kluwer Academic Publishers, Boston 2001.
11. Han J., Kamber M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco 2000.
12. Kłopotek M.:Inteligentne wyszukiwarki internetowe. EXIT. Warszawa 2001.

13. Murray M. Claffy K.: Measuring the Immeasurable: Global Internet Measurement Infrastructure. Proc. of PAM'2001, 2001.
14. Muraszkiewicz M.: Eksploracja danych dla telekomunikacji. PLOUG'2000.
15. Padmanabhan V. N., Qiu L.: Network Tomography Using Passive End-to-End Measurements. DIMACS Workshop on Internet and WWW Measurement, 2002.
16. Palmer Ch. R., Siganos G., Faloutsos M., Faloutsos Ch., Gibbons P. B.: The connectivity and fault-telerance of the Internet topology, Proc. of NRDM 2001, 2001.
17. Saroiu S., Gummadi K. P., Gribble S. D. King: Estimating Latency between Arbitrary Internet End Hosts, Proc. of SIGCOMM IMW 2002.
18. Srikant R., Yang Y.: Mining Web Logs to Improve Website Organization. Proc. of WWW10 Conference, 2001.
19. Using Intelligent Miner for Data. V6 Rel. 1, IBM Redbooks 1999, SH12-6394-00.
20. Wang M., Madhyastha T., Chan N.H., Papadimitriou S., Faloutsos C.: Data Mining Meets Performance Evaluation: Fast Algorithms for Modeling Bursty Traffic. Proc. 18th Int. Conference on Data Engineering, 2002.
21. Wolski R.:Dynamically Forecasting Network Performance Using the Network Weather Service. Technical Report TR-CS96-494, U.C. San Diego 1996.

Recenzent: Dr inż. Marcin Gorawski

Wpłynęło do Redakcji 14 kwietnia 2003 r.

Abstract

Modern Internet should be evaluated and analyzed from different point of views. This paper proposes the use of mining ideas for the analysis of performance data concerning Internet communication network. It is showed how particular data mining functions of the IBM Intelligent Miner for Data may be applied to the analysis of Internet problems.

Adres

Leszek BORZEMSKI: Politechnika Wrocławska, Instytut Sterowania i Techniki Systemów, ul. Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Polska, leszek@ists.pwr.wroc.pl.