Zastosowanie robota sieciowego w analizie internetowych systemów aukcyjnych

Artur Strzelecki, Tomasz Bacewicz, Marcin Ściański

Słowa kluczowe: **roboty sieciowe, aukcje internetowe**

Dynamiczny rozwój Internetu niesie ze sobą nowe możliwości, ale także nowe zagrożenia i wymagania. Oczywiste jest, że aby przemysł rozwijał się w sposób prawidłowy, dynamiczny i konkurencyjny wymagana jest jego jak najszersza automatyzacja. Korzystając dziś z najnowszych technologii pomiarowych automatyzujemy niemal każdy proces produkcyjny. Dzięki temu wyraźnie zwiększamy jego efektywność, poprzez obniżenie kosztów pracy ludzkiej oraz wyeliminowania błędów jakie człowiek może popełnić w procesie decyzyjnym. Podobnie należy postąpić w przypadku analizy tak ogromnego zasobu informacji jakim jest sieć Internet i wszystkich informacji zawartych na poszczególnych stronach. Dodanie dużych zasobów informacji do Internetu lub ich pobranie i analiza wymagają pełnej automatyzacji. Manualne zebranie i opracowanie takiego ogromu informacji jest niemożliwe do wykonania w czasie rzeczywistym. Mechanizmy automatycznie przeszukujące Internet w celu zebrania informacji wymaganych przez ich autora nazywamy robotami sieciowymi.

Roboty sieciowe

Najbardziej powszechnym rodzajem tego typu aplikacji są roboty indeksujące wyszukiwarek internetowych. Doskonałym przykładem może być Googlebot pracujący dla najpopularniejszej wyszukiwarki internetowej Google, który przechodzi pomiędzy poszczególnymi stronami korzystając z odnośników na nich umieszczonych. Tworzy on sieć powiązań pomiędzy stronami i na podstawie ich siły reprezentowanej wynikiem algorytmu

PageRank i jego późniejszych modyfikacji przedstawia rezultaty na stronie z wynikami wyszukiwania. Googlebot został zaprogramowany w dwóch wersjach. Pierwsza z nich to deepbot, która ma za zadanie zindeksować jak najwięcej informacji na użytek wyszukiwarki, drugi natomiast to freshbot, mający na celu odnalezienie w sieci nowo powstałych witryn internetowych, które przez pewien okres czasu są w wyszukiwarce Google promowane. Robotami tego typu możemy świadomie kierować wykorzystując plik robots.txt, w którym zawieramy instrukcje zezwalające na indeksowanie wybranych części serwisu i blokowanie innych. Roboty wyszukiwarek są aplikacjami użytecznymi dla całej internetowej społeczności. Ułatwiają nam szybkie odnalezienie wymaganych informacji. Niestety nie robią tego bezkosztowo, ponieważ wykorzystują zasoby transferu jakim dysponuje strona. Omówiony rodzaj programu jest robotem pobierającym informacje z sieci, ale to tylko jeden z wielu typów robotów sieciowych.

Drugą bardzo dużą i niestety szkodliwą dla twórców stron internetowych oraz administratorów for internetowych grupą robotów są spamboty. Spamboty mają za zadanie przede wszystkim umieszczanie i rozpowszechnianie informacji w sieci. Dwa najpowszechniejsze typy spam botów to ‘email spambot’ i ‘forum spambot’. Pierwszy z nich przeszukuje Internet w poszukiwaniu adresów e-mail, które zostają zapisane w jego bazie danych. Wyszukiwanie zostaje przeprowadzone w tekście reprezentującym kod strony internetowej pobranym poprzez protokół HTTP, za pomocą wyrażeń regularnych które opisują ogólny wzór składni poprawnego adresu e-mail np.: *\b[A-Z0-9._%+-]+@[A-Z0-9.-]+\.[A-Z]{2,4}\b*. Następnie na zebrane adresy e-mail rozesłana zostaje wiadomość najczęściej zawierająca treści reklamowe lub niebezpieczne odnośniki prowadzące do zainstalowania niepożądanego oprogramowania na naszym komputerze. Problemy jakie wiążą się z takimi programami to tworzenie niepotrzebnego

ruchu w Internecie, obciążającego jego infrastrukturę oraz zasypanie naszych skrzynek mailowych niechcianymi wiadomościami. Forum spambot służy natomiast do umieszczania informacji reklamowych na forach dyskusyjnych. Roboty tego typu są w stanie przejść poprzez proces rejestracji użytkownika i w losowym temacie założonym uprzednio na forum dodawać swoje wiadomości. Zawierają one najczęściej teksty reklamowe wraz z odnośnikami do promowanych stron. Masowe umieszczanie odnośników do wybranej strony na wielu forach prowadzi do wzrostu wartości PageRank w wyszukiwarce Google, pozwalając wypromować w ten sposób stronę internetową, skutkując jednocześnie przekłamaniem realnej wartości strony i trafności wyszukiwania. Programami typu spambot nie możemy w żaden sposób kierować. Wpływ jaki ma na nie użytkownik to próba ich zablokowania, poprzez zablokowanie adresów e-mail, z których wysyłane są wiadomości lub tych z których poszczególne roboty rejestrują się jako użytkownicy forum. Dodatkowo w procesie rejestracji wymaga się odczytania grafiki zawierającej ciąg znaków zapisany w sposób utrudniający podejrzenie go za pomocą systemów rozpoznających znaki. W sieci istnieją również strony tworzące masowe ilości losowych (nieistniejących) adresów e-mail, zawierających jednocześnie odnośniki do swoich kopii, skutecznie zapełniając bazę danych robotów bezużytecznymi informacjami.

Najbardziej neutralnym typem robotów sieciowych są programy tworzące kopie stron na użytek off-line lub pomagające w przeniesieniu ich na serwer zapasowy. Kopiują one informacje jedynie w ramach jednej domeny nie przechodząc i nie pobierając danych z linków zewnętrznych. Aplikacje tego typu najpowszechniej stosowane były w czasach gdy dostęp do łączy szerokopasmowych i stałego dostępu do Internetu był silnie ograniczony.

Przyglądając się sposobowi przepływu danych w globalnej sieci oraz wykorzystaniu robotów sieciowych możemy intuicyjnie dokonać pewnego porównania Internetu do wielkiego rynku informacji, na którym pewne zbiory cennych danych są strzeżone a dostęp do nich zostaje ograniczony lub staje się odpłatny. Natomiast inne informacje za pomocą organizacji marketingowych jakimi niewątpliwie stają się roboty sieciowe zostają wypromowane. Co więcej za pomocą robotów sieciowych jesteśmy w stanie zebrać interesujące nas informacje na temat wybranej grupy społecznej i skutecznie wykorzystać je w celu promowania lub rozwoju naszej strony internetowej. Należy pamiętać jednak, że ocena sposobu owej promocji oraz konkurencji zawsze styka się z tematem etyki działania w Internecie. Budując aplikacje sieciowe nie zapominajmy o tym, aby ich działanie nie było uciążliwe dla współużytkowników Internetu. Nie należy naruszać ich prawa do prywatności, a jeżeli sami posiadamy ważne informacje powinniśmy zadbać o ich poprawne zabezpieczenie.

Obszary działania robota sieciowego

Pierwszym napisanym i wykorzystanym przez autorów robotem sieciowym był program służący do pobierania i zapisywania wybranych i powszechnie dostępnych informacji na temat transakcji dokonywanych na portalu aukcyjnym Allegro. Zamierzeniem autorów była analiza komentarzy wystawianych po zakończeniu aukcji internetowych i określenie na tej podstawie poziomu uczciwości poszczególnych sprzedawców. Realizacja zaplanowanych funkcjonalności wymagała sprzęgnięcia robota z bazą danych MySQL. Pozwoliło to zapisywać napływające dane w czasie rzeczywistym i jednocześnie je porządkować. Wystarczająca okazała się baza danych składająca się z dwóch tabel. W pierwszej mieściło się 6 kolumn, w drugiej tylko 3. Autorzy w pierwszej tabeli zapisywali kolejno:

- identyfikator transakcji, który przyjmował unikatowe wartości, począwszy od jedności,
- identyfikator sprzedawcy,
- identyfikator kupującego,
- numer aukcji,
- typ komentarza, przyjmujący wartość 2 dla komentarza pozytywnego, wartość 1 dla komentarza neutralnego oraz wartość 0 dla negatywnie ocenionej transakcji,
- datę wystawienia komentarza.

W drugiej tabeli pojawiały się następujące informacje:

- numer użytkownika Allegro,
- nazwa użytkownika,
- stan konta (zawieszony / aktywny).

Sam program napisano w języku Perl, przy wykorzystaniu bibliotek LWP::UserAgent oraz LWP::RobotUA. Początkowo autorzy przygotowali wersję działającą w oparciu o język PHP i bibliotekę cURL, jednak wersja ta, przy zastosowaniu identycznego algorytmu, okazała się znacznie wolniejsza. Ponadto serwer Allegro identyfikował robota jako przeglądarkę i po wysłaniu ogromnej ilości żądań z jednego komputera w sposób regularny, zostawało nakładane ograniczenie na szybkość połączenia, co w praktyce uniemożliwiało dalsze gromadzenie danych. Przejście na biblioteki języka Perl dało możliwość lepszej kontroli nad tym, jak identyfikowany jest skrypt w sieci oraz jakie ślady pozostawia na serwerach. To pozwoliło na nienaruszoną, kilkudniową pracę skryptu. Nadanie robotowi „tożsamości”, inicjalizacja połączenia z bazą danych oraz ustawienie częstotliwości wysyłania zapytań do serwera odbywały się w pierwszych dwudziestu pięciu liniach kodu.

Rdzeniem skryptu była sekcja, składająca się z dwóch zagnieżdżonych pętli, w której następowała analiza zawartości strony internetowej portalu. W tej części wykorzystano mechanizm wyrażeń regularnych, wydajnie działający w języku Perl. Właściwość ta sprawia, że język Perl jest bardzo popularnym i co najważniejsze, efektywnym narzędziem do pracy z tekstem. Poprzez porównywanie zawartości pobranej strony internetowej ze specjalnie przygotowanymi wzorcami, robot był zdolny do pozyskania tylko niezbędnych treści oraz zapisania ich w odpowiednich zmiennych. Takie działanie przyniosło wiele korzyści, gdyż początkowo program zapisywał treść całej strony do pliku na dysku, a dopiero potem następowało parsowanie tekstu. Zmiana podejścia przyniosła znaczną oszczędność ograniczonej przestrzeni dyskowej oraz wyraźnie zauważalny wzrost szybkości działania programu. Ostatnim etapem było zapisanie w bazie danych wartości przechowywanych w zmiennych.

Program sprawdzał, czy pobranie danej strony powiodło się, a w razie niepowodzenia ponownie ją pobierał. Dodatkowo robot sprawdzał czy dany użytkownik istnieje lub czy jego konto nie zostało zablokowane. Całość zajmowała około 140 linii kodu, które umożliwiły pobranie informacji o transakcjach około 7 milionów użytkowników.

W oparciu o pierwsze doświadczenia z językiem Perl autorzy przygotowali kolejny program. Tym razem robot przeszukiwał portal społecznościowy nasza-klasa.pl. Zamierzeniem autorów było sprawdzenie, czy możliwe jest masowe pobranie danych osobowych użytkowników portalu. Pomysł zrodził się w związku z kontrowersjami wokół bezpieczeństwa danych przechowywanych na serwerach portalu. Poruszono także problem legalności publikowanych informacji w świetle ustawy o ochronie danych osobowych. Jak więc widać, celem robota było

zgromadzenie możliwie jak największej ilości danych, pomimo zabezpieczeń wprowadzonych w portalu.

Sam mechanizm pobierania informacji działał identycznie jak w pierwszym programie. Odpowiednio przygotowane wyrażenia regularne i połączenie z bazą danych MySQL pozwoliły skutecznie gromadzić dane takie jak: imię, nazwisko, pseudonim szkolny, nazwisko rodowe (jeśli występowało) telefon, miasto, wiek, płeć, numer w komunikatorze Gadu-gadu oraz nazwę użytkownika w komunikatorze Skype.

Jednakże aby uzyskać dostęp do poszukiwanych informacji trzeba posiadać w portalu konto i przejść proces logowania. Z pomocą przyszła kolejna biblioteka Perla WWW::Mechanize, która umożliwiała wypełnienie odpowiednich pól w formularzu na stronie i przesłaniu danych. Pozostawało tylko stworzyć fikcyjne konto. Niestety, po pobraniu określonej liczby stron z danymi kolejnych użytkowników mechanizmy zabezpieczające portal blokowały tymczasowo możliwość przeglądania witryny naszemu nowo stworzonemu użytkownikowi.

W trakcie kilku prób uruchomienia robota okazało się, że portal nasza-klasa.pl nie pozwala automatycznie przeglądać większej liczby stron jednemu użytkownikowi. Podjęto próby zbliżenia zachowania robota do bardziej naturalnego dla zwykłych użytkowników (wyciete user-agent i referer). Niestety wydłużyło to tylko nieznacznie moment, w którym nasz wirtualny użytkownik o nazwie „pająk informatyczny” był blokowany przez portal. W tym krótkim czasie około 15 minut robot był w stanie pobrać na łączu o prędkości 512 kbps informacje o jednym tysiącu użytkowników. Nie ulega wątpliwości, że to działanie zabezpieczające zostało podjęte z uwagi na liczne już próby pobrania danych z portalu. Z pewnością świadczy o tym również komunikat

jaki zostaje wyświetlony po zablokowaniu konta z powodu nadmiaru pobranych danych: „Serwery są przeciążone, prosimy spróbować ponownie…”.(WYCIETY RYSUNEK Z PANEM GABKA)

Aby obejść to zabezpieczenie autorzy zdecydowali się na wykorzystanie wielopoziomowych serwerów proxy, co dało możliwość regularnej zmiany adresu IP komputera z którego wysyłano żądania. Takie rozwiązanie choć skuteczne, znacznie spowolniło proces pobierania danych i było o wiele mniej stabilne – zdarzało się ze robot tracił połączenie z serwerami proxy.

Do bazy została pobrana próbka danych wielkości 5022 rekordów. W tej grupie ok. 10% użytkowników podało swój numer komunikatora Gadu-gadu, jednak już tylko mniej niż 2% użytkowników podało swoją nazwę w komunikatorze Skype. Numer telefonu był dostępny u około 2% użytkowników. Powyższe wyliczenia pokazują, że użytkownicy pragnąc zachować pewien określony stopień prywatności, nie podają informacji umożliwiających bardzo łatwy, bezpośredni kontakt. Pomimo euforii panującej wokół portalu, użytkownicy zdają się rozumieć związane z nim zagrożenia.

Analiza transakcji w systemie aukcyjnym

Stworzony mechanizm robota sieciowego pozyskuje pełną informację o wszystkich zrealizowanych transakcjach aukcyjnych oraz o uczestnikach tych transakcji. Pełna kolekcja danych trwała 9 dni i 3 godziny, kosztowała około 510GB transferu i zebrała wszystkie komentarze o zakończonych aukcjach od początku istnienia serwisu do marca 2008. Badaniom zostały poddane następujące informacje o transakcjach na aukcjach internetowych:

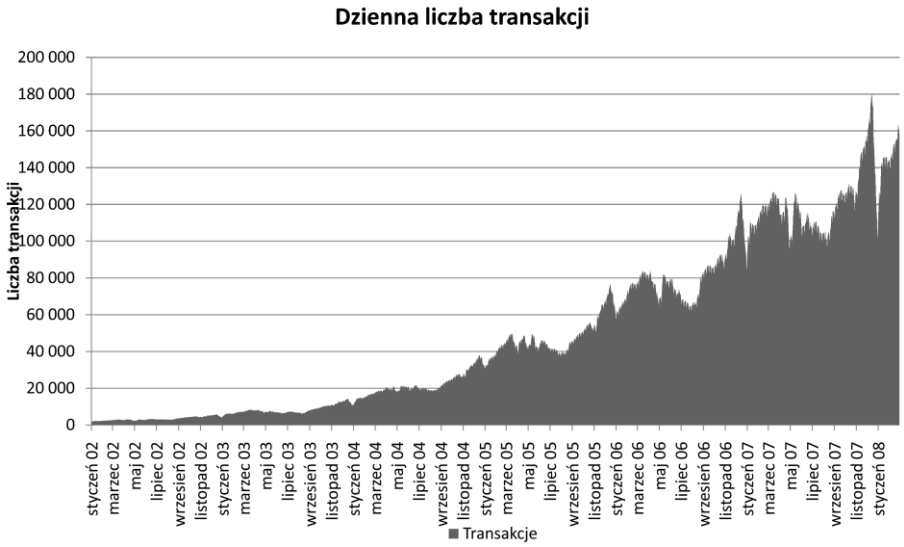
Tabela 1.

Statystyka zebranej kolekcji danych

Liczba transakcji	112498829
Komentarze pozytywne	111169681
Komentarze neutralne	513982
Komentarze negatywne	815166
Liczba sprzedających	1686954
Liczba kupujących	3612891

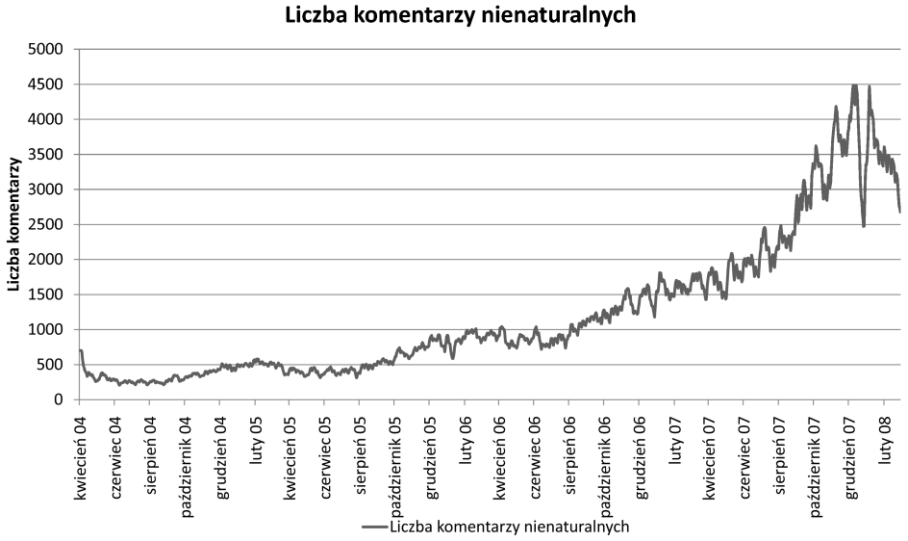
Źródło: opracowanie własne

Rysunek 1 ilustruje liczbę transakcji zakończonych wystawieniem komentarza, na rysunku nie ma transakcji z lat 2000-2001, których jest znikoma liczba. Wraz z upływem czasu widać wzrost liczby realizowanych transakcji. Dynamika wzrostu utrzymuje się na średnim poziomie 109% rocznie, przy czym w ostatnim roku było to tylko 49%. Wyniki pokazują rosnące znaczenie zawieranych transakcji C2C. Grudzień i marzec każdego roku cechują się silnym skokiem w górę, co związane jest ze wzmożonymi zakupami świątecznymi.



Rys 1. Dzienna liczba transakcji w latach 2002-2007

W trakcie empirycznej analizy danych autorzy wykryli bardzo niepokojącą sytuację. Przy zawieszonych kontach użytkowników z pierwszej dziesiątki najgorszych sprzedawców cały czas pojawiają się pozytywne komentarze. Chronologiczny numer aukcji ujawnia, że są to aukcje, które odbyły się na kilka miesięcy do kilku lat przed zablokowaniem konta sprzedawcy. Prawdopodobnie istotna część komentarzy jest wystawiana przez uczestników posiadających fałszywe tożsamości i posługujących się nimi do manipulowania opinią sprzedawców. Przy kontach użytkowników, którzy nie są zawieszeni przez serwis, a posiadają ogromną liczbę przeprowadzonych transakcji, spora część komentarzy jest również wystawiana do aukcji, które odbyły się w okresie od kilku miesięcy do kilku lat wstecz. Rysunek 2 przedstawia liczbę wystawionych komentarzy do transakcji z co najmniej sześciomiesięcznym opóźnieniem od rzeczywistego końca transakcji. Takich komentarzy jest ponad 1,5% łącznej liczby transakcji. Niepokoi to, że zachowanie dotyczy sprzedających z największym wolumenem sprzedaży.

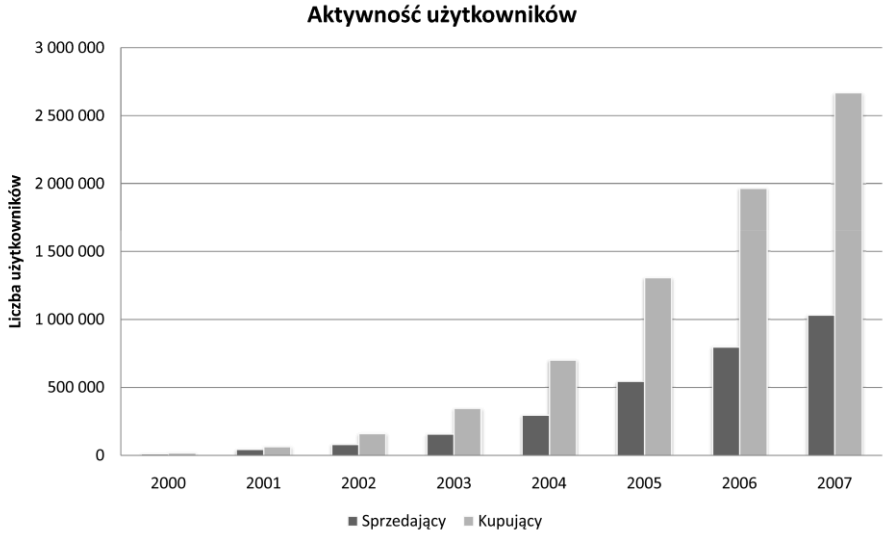


Rys 2. Liczba transakcji, do których komentarz jest opóźniony co najmniej o 6 miesięcy

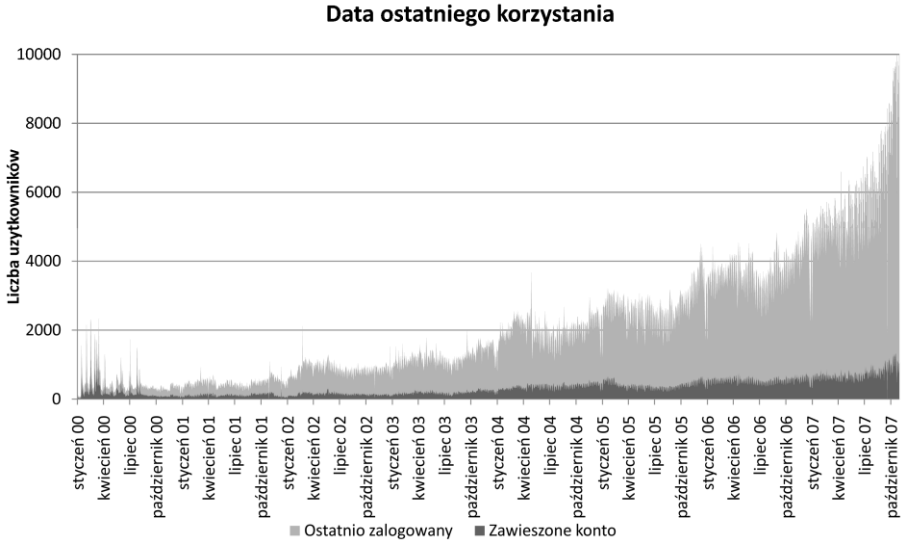
Rysunek 3 przedstawia dynamikę przyrostu użytkowników serwisu Allegro wraz z jego zagranicznymi oddziałami w Bułgarii, Czechach, Rosji, Rumunii, Słowacji, Ukrainie i na Węgrzech. Zaobserwowano stałą tendencję wzrostową, osiągającą od początku 2008 roku średni przyrost 7921 użytkowników dziennie. Co ciekawe, na wykresie widać, że liczba nowo rejestrowanych użytkowników posiada minima lokalne pod koniec grudnia oraz przez cały lipiec każdego roku. Pobrano informacje o 10 054 530 kontach użytkowników. 12% wszystkich kont jest zawieszonych, natomiast 26% kont nigdy nie było używanych.

Rys 3. Dzienna liczba rejestracji w latach 2000-2007

Rysunek 4 pokazuje dynamikę aktywności użytkowników serwisu z podziałem na pełnione funkcje sprzedających i kupujących. Na wykresie znajduje się łączna liczba kupujących i sprzedających biorących udział w transakcjach w rocznym przekroju. Z każdym rokiem zwiększa się liczba kontrahentów biorących udział w transakcjach.



Rys 4. Aktywność użytkowników z podziałem na pełnione role w kolejnych latach



Rys 5. Data ostatniego zalogowania się użytkownika do serwisu, ciemniejszy obszar dotyczy użytkowników z zawieszonym kontem

Rysunek 5 pokazuje ostatnią datę zalogowania się do systemu Allegro wszystkich jego użytkowników do końca października 2007 roku

uwzględniając użytkowników z zawieszonymi kontami. Dotyczy to tylko tych użytkowników, którzy kiedykolwiek byli aktywni na Allegro, czyli co najmniej raz zalogowali się do serwisu po założeniu konta. Od początku roku do 10 marca br. z Allegro aktywnie korzystało 3546640 użytkowników. Z wykresu wykluczono dwie daty, w których serwis Allegro dokonywał strategicznych decyzji rozwojowych. Dotyczy to zmiany rosyjskiej nazwy swojego odpowiednika z Aukro.ru na Motolok.ru, w ten dzień przestano monitorować ponad 700 tys rosyjskich użytkowników oraz wycofanie się z rynku holenderskiego, gdzie korzystało ponad 120 tysięcy użytkowników. Na wykresie niższa warstwa reprezentuje użytkowników z zawieszonymi kontami.

Podsumowanie

Analiza poszczególnych parametrów w określonych przedziałach czasowych pozwala ukazać, w jaki sposób zmienia się wolumen transakcji oraz sposób zachowania sprzedawców. Ponadto wskazuje na obecność w serwisie grupy komentarzy „nienaturalnych" tzn. nadmiernie opóźnionych do daty zawarcia transakcji oraz dużej ilości komentarzy pozytywnych pojawiających się na koncie użytkownika, długo po jego zawieszeniu.

Istotne jest, że cały proces analizy i wnioskowania został przeprowadzony w oparciu o dane największej liczby transakcji i użytkowników ich realizujących, jaka do tej pory brała udział od początku istnienia serwisu Allegro. Jesteśmy przekonani ze dogłębna analiza posiadanych informacji rozszerza naszą wiedzę o wzorcach zachowań użytkowników serwisów aukcyjnych, dzięki której możemy z większym zaufaniem podejść do internetowych transakcji aukcyjnych.

Literatura

1. A. Strzelecki, T. Bacewicz, M. Ściański, Bezpieczeństwo danych w internetowych sieciach społecznych na przykładzie portalu nasza-klasa.pl, [w:] Systemy Wspomagania Organizacji SWO 2008, pod

red. Małgorzaty Pańkowskiej T. Porębskiej-Miąc i H. Sroki, Wydawnictwo Akademii Ekonomicznej, Katowice 2008, CrossRef
2. A. Strzelecki, T.Bacewicz, M.Ściański, Ocena zachowania użytkowników platformy handlu C2C na podstawie eksploracji danych i ich aktywności w internetowym systemie aukcyjnym, e-mentor 2008, numer 2(24), s.76-82, CrossRef

3. A. Strzelecki, Transakcje na aukcjach internetowych źródłem wiedzy o jakości usług sprzedawców, Rozdział w monografii: Bazy danych: Rozwój metod i technologii, Praca zbiorowa pod redakcją: S. Kozielskiego, B. Małysiak, P. Kasprowskiego, D. Mrozka, Wydawnictwa Komunikacji i Łączności, Warszawa 2008, s. 175-184, CrossRef