

# Analiza wzorców nielegalnych zachowań w aukcjach internetowych

Jakub Piasecki<sup>1</sup>, Wojciech Roczniak<sup>1</sup>,  
Anna Zygmunt<sup>1</sup>, Jarosław Koźlak<sup>1</sup>

**Streszczenie:** Transakcje internetowe stanowią coraz większą część ogólnego obrotu handlowego. Istotną rolę w handlu internetowym odgrywają aukcje. Popularność aukcji przyciągnęła niestety oszustów, którzy żerują na uczciwych użytkownikach. Niniejsza praca poświęcona jest problemowi wykrywania i przewidywania oszukańczych zachowań uczestników aukcji w internecie. Autorzy dokonują przeglądu przestępstw popełnianych na serwisach aukcyjnych oraz dotychczasowych publikacji dotyczących ich wykrywania. Następnie, przedstawiają własne mechanizmy kolekcjonowania danych o aukcjach z różnych źródeł oraz przechowywania ich w strukturze ontologicznej. Oceniają skuteczność dotychczasowych metod, a także przeprowadzają analizę danych, w wyniku której wytypowane zostały nowe współczynniki zwiększające skuteczność algorytmów. Stawiane tezy poparte są serią eksperymentów.

**Słowa kluczowe:** aukcje internetowe, oszustwa internetowe, ontologia, klasyfikacja

## 1. Wstęp

Według raportu *Internet 2006* przygotowanego przez Gemius SA (Gemius SA (2006)) obecnie w Polsce z Internetu korzysta ponad 12 milionów osób<sup>2</sup>. W ciągu roku 2006 zakupy w Internecie dokonało 55% z nich, co stanowi wzrost o jedną trzecią w stosunku do roku 2005. Najbardziej popularną formą handlu elektronicznego są aukcje internetowe, na których kupowało 77% internautów. Użytkownicy aukcji najbardziej cenią sobie: niskie ceny, duży wybór oraz wygodę kupowania o dowolnej porze.

Przytoczone powyżej dane statystyczne pokazują, jak ogromnym zainteresowaniem cieszą się w Polsce aukcje internetowe. Co ważniejsze, badania wskazują stabilną tendencję wzrostową (Money.pl). Biorąc pod uwagę to, że pod względem ilości internautów w stosunku do liczby mieszkańców w Polsce ciągle wiele brakuje do innych krajów, można śmiało prognozować dalszy rozwój handlu elektronicznego w Polsce.

Na polskim rynku aukcji internetowych pierwsze miejsce zajmuje zdecydowanie serwis Allegro.pl<sup>3</sup>, który według raportu SMG/KRC (SMG/KRC) ma około

<sup>1</sup> Katedra Informatyki, AGH, Al. Mickiewicza 30, 30-059 Kraków

e-mail: {piasecki,roczniak}@student.agh.edu.pl, {azygmunt,kozlak}@agh.edu.pl

<sup>2</sup> stan z września 2006

<sup>3</sup> <http://www.allegro.pl>

6,5 miliona zarejestrowanych użytkowników. Na drugim miejscu znajduje się polski oddział eBay<sup>4</sup> z dwoma milionami użytkowników. Kolejne miejsca zajmują: Świstak.pl<sup>5</sup> oraz serwis aukcje24.pl<sup>6</sup>. Poza wymienionymi powyżej istnieje kilka innych serwisów, których liczba użytkowników nie przekracza kilkudziesięciu tysięcy. Niestety, duży obrót pieniędzy generowany poprzez handel internetowy przyciąga wszelkiego rodzaju oszustów. Przepęstwa popełniane na aukcjach internetowych możemy podzielić na następujące kategorie<sup>7 89</sup>:

- Wysłanie towaru niezgodnego z opisem.
- Niedostarczenie towaru po otrzymaniu pieniędzy.
- Dostarczenie przedmiotu niezgodnego z opisem, lub niepoinformowanie kupującego o istotnych cechach np. uszkodzeniu.
- Oferowanie za pośrednictwem serwisów aukcyjnych towarów pochodzących z przestępstwa.
- Wycofanie płatności. Oszust przedstawia się jako osoba z zagranicy, i oferuje kupno przedmiotu po wysokiej cenie. Dokonuje wpłaty na konto sprzedającego, a następnie wykorzystuje możliwość cofnięcia transakcji, jaką mają posiadacze niektórych kart kredytowych.
- Fałszywe strony typu Escrow. Oszust po wygraniu aukcji, symuluje wpłatę pieniędzy z wykorzystaniem strony typu Escrow, która w rzeczywistości została zarejestrowana przez niego z użyciem kradzionego numeru karty internetowej. Użytkownik po zauważeniu informacji o wpłacie na stronie, wysłał towar (patrz Steiner 2005).

Można też wyróżnić całą grupę przestępstw opierających się na kradzieży tożsamości.

Poza wyżej wymienionymi należy wspomnieć o nadużyciach związanych ze strukturą aukcji internetowej i systemem komentarzy: Może ono polegać na nieuczciwym budowaniu reputacji. Użytkownik wykorzystuje nieuczciwe sposoby zebrania jak największej ilości pozytywnych komentarzy (np. wystawianych przez znajomych lub dzięki zakupowi dużej liczby tanich przedmiotów). Inną możliwością nadużyć jest sztuczne podbijanie ceny przedmiotu aukcji. Użytkownik korzystając z drugiego konta lub z pomocą współnika podbija cenę oferowanego przez siebie przedmiotu. Nieuczciwy sprzedający może także namawiać kupującego do zawarcia transakcji poza systemem aukcyjnym oferując mu niższą cenę. Kupujący godząc się na to traci całą ochronę jaką może mu zagwarantować serwis aukcyjny np. ubezpieczenie.

<sup>4</sup> <http://www.ebay.pl/>

<sup>5</sup> <http://www.swistak.pl>

<sup>6</sup> <http://aukcje24.pl>

<sup>7</sup> <http://www.policja.pl>

<sup>8</sup> [http://www.pueblo.gsa.gov/cic\\_text/computers/internet-auction/internet\\_auctions.htm](http://www.pueblo.gsa.gov/cic_text/computers/internet-auction/internet_auctions.htm)

<sup>9</sup> <http://www.crime-research.org/articles/Wahab1>

## 2. Przegląd stosowanych metod

W stosunku do ogromnej popularności handlu elektronicznego, a w szczególności aukcji internetowych, ilość publikacji naukowych, których przedmiotem jest badanie przestępstw z tej dziedziny, jest ograniczona.

Najbardziej zaawansowane podejście przedstawione jest w artykułach opartych na pracach prowadzonych na Carnegie Mellon University, gdzie prezentowane są metody wykrywania zorganizowanych działań przestępczych polegających na niedostarczeniu towaru przez sprzedającego. W pierwszym artykule (Chau (2005)), prezentowane są metody uczenia maszynowego pozwalające na klasyfikację podejrzanych zachowań na podstawie informacji o użytkownikach, którzy popełnili przestępstwo. Autorzy typują parametry charakteryzujące historię działalności użytkownika na aukcjach internetowych, a następnie przeprowadzają ocenę stworzonego przez siebie modelu. W artykule drugim (Chau (2005a)), podjęto temat współpracy przestępców podczas dokonywania oszustwa. Prezentowany jest algorytm *2-Level Fraud Spotting*, który w pierwszej fazie typuje przestępców w sposób opisany w publikacji pierwszej, a następnie bada sieć interakcji pomiędzy użytkownikami w poszukiwaniu podejrzanych powiązań. Publikacja trzecia (Chau (2007)) dotyczy podejścia do budowy dużego, skalowalnego systemu *NetProbe* w oparciu o wcześniejsze prace. Prezentowana jest także wersja iteracyjna algorytmu, likwidująca konieczność przebudowywania całego modelu po pobraniu nowych danych. Oprócz wyżej wymienionych publikacji należy wymienić serię artykułów dotyczących budowania reputacji na serwisach aukcyjnych i nadużyć w tym zakresie np. pracę (Wang (2005)). Pośrednio związany jest także artykuł dotyczący fikcyjnych płatności kartą kredytową na aukcjach internetowych (Chae (2007)), gdzie autorzy w oparciu o analizę przebiegu aukcji starają się wykryć podejrzane zachowania.

## 3. Koncepcja systemu do wykrywania oszukańczych zachowań

Głównym zadaniem jakie postawiliśmy przed naszym systemem jest gromadzenie oraz analiza danych o aukcjach i użytkownikach serwisu aukcyjnego Allegro. Jednym z celów tych analiz jest wyznaczenie wzorca zachowania oszusta aukcyjnego, oraz określenia na ile możliwe jest jego wykrycie zanim zacznie oszukiwać. Możemy wyróżnić dwa główne obszary aktywności systemu: gromadzenie danych oraz ekstrakcja cech i ich analiza.

### 3.1. Gromadzenie danych

Dane opisujące aukcje lub użytkowników dostępne są w postaci tekstu na stronie internetowej. Jeden z modułów naszego programu zajmuje się wyodrębnieniem tych danych i udostępnieniem modułowi odpowiedzialnemu za zapisanie ich w ontologii. System przystosowany jest do współpracy nie tylko z serwisem *Allegro*, ale również z innymi źródłami danych. System *Allegro* nie udostępnia danych historycznych, co stanowi niemały problem w przypadku analizy dłuższej działających użytkowników. Dlatego też rozszerzyliśmy nasz system o możliwość pobierania danych z alterna-

tywnego źródła. W przypadku braku danych w systemie Allegro korzystamy z alternatywnego źródła jakim jest serwis archiwizujący aukcje *Allegro Archiver*<sup>10</sup>, co umożliwia uzupełnienie brakujących danych. Zgromadzone dane na temat aukcji i użytkowników pozwalają nam wyznaczyć różnego rodzaju cechy opisujące ich zachowanie. Te wartości służą potem do przeprowadzania analiz i określania zachowania użytkownika. Budowa systemu pozwala również na przeprowadzenie wielu różnych eksperymentów, które korzystają ze zgromadzonych danych.

### 3.2. Rola ontologii

Podczas budowy systemu jednym z problemów był brak znajomości kompletnego zbioru danych, które potrzebne są do zbudowania optymalnego modelu. Spośród wszystkich danych dostępnych na serwisach aukcyjnych, należało wybrać tylko te, które są niezbędne. Pobieranie nadmiarowych danych wiązało się z narzutami czasowymi, co przy tak dużej ich ilości nie było bez znaczenia. Z drugiej strony zależało nam na zachowaniu jak najbardziej elastycznej struktury, która umożliwiłaby jej późniejsze zmiany bez potrzeby ponownego pobierania danych. Podejście to miało nam zapewnić jak największą swobodę podczas przeprowadzania eksperymentów.

Serwis internetowy *Allegro* ogranicza przeglądanie historii aukcji użytkowników do kilku miesięcy wstecz. W celu zbudowania jak najpełniejszego zbioru uczącego skorzystaliśmy z uprzejmości serwisu *AllegroArchiver*, który archiwizuje informacje o aukcjach. Potrzeba pobierania danych z heterogenicznych źródeł zmusiła nas do zaprojektowania struktury, która w najprostszy możliwy sposób umożliwiałaby ich integrację. W miarę możliwości struktura ta miała zapewnić możliwość rozbudowy systemu dla innych serwisów aukcyjnych np. *eBay*<sup>11</sup>. Powyższe założenia zostały całkowicie spełnione poprzez przechowywanie pobranych danych w formacie języka opisu ontologii *OWL*. Ontologia zapewnia elastyczną rozbudowę, dzięki temu, że dane jak i ich opis zapisane są w tym samym języku. Dzięki temu możemy dodać np. własność do klasy nawet kiedy ta posiada już instancje. Przy wyborze ontologii kierowaliśmy się również dostępnością narzędzi i ich przydatnością w analizie danych. Do tworzenia ontologii korzystaliśmy z narzędzia *Protege*<sup>12</sup>, które umożliwia podłączenie silnika wnioskującego, a także zadawanie zapytań w języku *SPARQL*<sup>13</sup>. Dostępne narzędzia do wizualizacji takie jak *RDFGravity*<sup>14</sup> ułatwiają wizualizację danych bez dodatkowej transformacji, co wpływa na wygodę analizy. Należy jednak nadmienić, że obecnie większość z narzędzi do wizualizacji umożliwia pracę jedynie ze stosunkowo małymi zbiorami danych.

W czasie realizacji systemu okazało się, że wybór ontologii był słuszny również z innych powodów. Mianowicie, struktura ontologii w prosty i logiczny sposób opisuje dziedzinę problemu w postaci grafowej. Struktura grafu odwzorowuje w sposób naturalny interakcje zachodzące pomiędzy uczestnikami aukcji. Dodatkowo ontologia *OWL DL* pozwala na dodawanie reguł na wzór systemów ekspertowych,

<sup>10</sup> <http://www.allegro.archiver.pl/>

<sup>11</sup> <http://www.ebay.pl/>

<sup>12</sup> <http://protege.stanford.edu/>

<sup>13</sup> <http://www.w3.org/TR/rdf-sparql-query/>

<sup>14</sup> <http://semweb.salzburgresearch.at/apps/rdf-gravity/index.html>

co umożliwiła rozbudowę systemu w oparciu o dotychczas dostępną wiedzę.

## 4. Budowa systemu

### 4.1. Schemat architektury

W trakcie zbierania wymagań i analizy systemu staraliśmy się wyodrębnić funkcjonalne części i zapewnić modułarną budowę. Dzięki temu prace nad systemem i jego konserwacja mogły przebiegać równolegle i nie wymagały dużych nakładów pracy. Stworzony system składa się z kilku niezależnych modułów odpowiadających za realizację wyodrębnionych funkcji systemu takich jak:

- **Parsowanie danych** - parser udostępnia innym modułom strony aukcji internetowych w postaci poprawnie sformatowanego dokumentu XML;
- **Ekstrakcja danych ze stron** - ekstrahuje dane dotyczące aukcji i użytkowników, a następnie udostępnia je innym modułom w postaci obiektów;
- **Pobierania danych** - algorytmy opisujące sposób pobierania danych;
- **Model danych** - dane oraz zależności pomiędzy nimi opisane językiem OWL;
- **Eksport danych** - odpowiada za eksport danych do różnych formatów;
- **Ekspertymenty** - ten moduł pozwala na przeprowadzanie eksperymentów na zgromadzonych danych, eksperymenty mogą być przeprowadzone zarówno na wyeksportowanych danych jak i na samej ontologii.

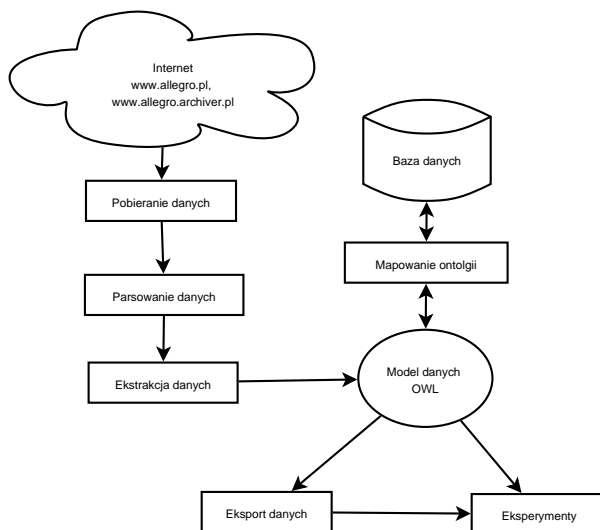
Schemat architektury systemu przedstawiony jest na rys.1.

### 4.2. Sposób reprezentacji danych

Pozyskane dane przechowywane są w bazie danych. Odbywa się to z użyciem odpowiednich narzędzi<sup>15</sup> umożliwiających mapowanie ontologii na relacyjną bazę danych. Takie podejście pozwoliło na wyeliminowanie konieczności operowania na dużych plikach do przechowywania ontologii i na zwiększenie wydajności systemu. Zbudowana została ontologia (rys. 2), której punktem centralnym są trzy koncepty, odpowiadające za reprezentację użytkowników serwisu aukcyjnego oraz za zachodzące między nimi interakcje: *User* (reprezentuje użytkownika serwisu aukcyjnego), *Auction* (reprezentuje pojedynczą aukcję), *Comment* (komentarz dobrowolnie wystawiany przez użytkowników po zakończeniu aukcji, wyrażający ich zadowolenie z przeprowadzonej transakcji).

Zdefiniowaliśmy także relacje odpowiadające za interakcje pomiędzy uczestnikami, które mają miejsce na aukcjach internetowych. Odwzorowujemy relacje zapoczątkowania aukcji, podbicia ceny, wygrania aukcji, a także wystawienia komentarza. Każdy koncept zawiera także specyficzne dla niego właściwości. Tak więc

<sup>15</sup> Jena, <http://jena.sourceforge.net/>



Rysunek 1. Schemat architektury systemu

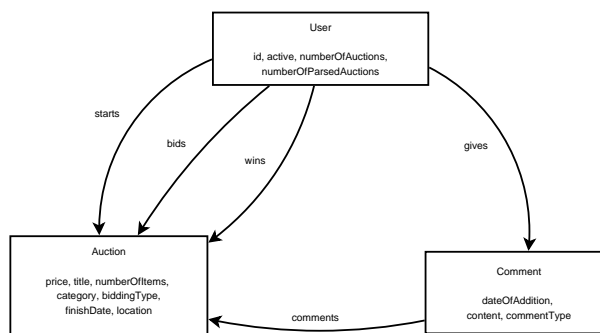
koncept *User* posiada własności takie jak: id, informacja o tym, czy użytkownik jest aktywny, czy jego konto zostało zablokowane, a także informacje pomocnicze użyteczne przy śledzeniu parsowania danych z Internetu. Koncept reprezentujący komentarz ma właściwości określające jego typ, datę wystawienia oraz treść. Właściwości konceptu *Auction*, przechowują informacje dotyczące szczegółów aukcji takich jak: cena wywoławcza, typ aukcji (zwykła lub *Kup Teraz*), data zakończenia, ilość wystawionych przedmiotów itp.

#### 4.3. Stosowane algorytmy

W aktualnej wersji ocena użytkowników i wskazywanie podejrzanych osób oparte jest na metodzie drzew decyzyjnych i wykorzystuje algorytm *C4.5*. Algorytm ten jest szeroko stosowany i ma kilka zalet, które szczególnie wpłynęły na jego wybór: może operować na atrybutach będących wartościami ciągłymi, umożliwia sklasyfikowanie rekordów z nieznaną wartością (cechę wykorzystywaną do określenia, czy dana osoba jest oszustem). Dodatkowo, otrzymane w podejściu drzew decyzyjnych reguły dają się łatwo wizualizować i są zrozumiałe dla człowieka. W dalszej części artykułu przedstawimy drzewa, które zostały wygenerowane z przygotowanych przez nas danych. Na wybór algorytmu wpłynęła także chęć porównania naszych wyników z wynikami zaprezentowanymi w (Chau (2005)), gdzie również stosowane było *C4.5*.

## 5. Wyniki

W ramach eksperymentów początkowo powtórzono eksperyment amerykański, który jednak nie wypadł pomyślnie w polskich realiach. Dlatego wprowadzono



Rysunek 2. Schemat ontologii aukcji internetowych

kilka nowych parametrów, które dużo lepiej charakteryzują przestępców w polskich warunkach.

### 5.1. Schemat eksperymentów

Pierwszą fazą było przygotowanie zbioru uczącego, który zawierałby reprezentatywne dane. Zbiór taki powinien zawierać oszustów oraz zwykłych użytkowników. W celu wyszukania profili oszustów przeszukaliśmy serwisy informacyjne pod kątem informacji dotyczących przestępstw na aukcjach. Niezwykle pomocna okazała się także strona *biskupi.pl*<sup>16</sup> prezentująca rankingi użytkowników *Allegro*. Skorzystaliśmy z rankingu prezentującego użytkowników z największą ilością negatywnych komentarzy. Przeglądając treść komentarzy ustaliliśmy, którzy z nich popełnili oszustwo polegające na niedostarczeniu towaru. Nasz zbiór uczący zawierał 27 przestępców i 282 uczciwych użytkowników. Kolejny krok polegał na stworzeniu modelu drzewa decyzyjnego i określeniu jego poprawności. Ostatnim krokiem eksperymentu było testowanie nowych użytkowników i sprawdzanie wyników klasyfikacji. Ten etap musieliśmy wykonywać ręcznie, a polegał on na sprawdzeniu aukcji wytypowanego użytkownika i ocenie na podstawie komentarzy czy dany użytkownik może być oszustem.

### 5.2. Powtórzenie amerykańskiego eksperymentu

Pierwszym eksperymentem była próba powtórzenia wyników jakie otrzymali autorzy artykułu (Chau (2005)) w polskich realiach. W tym celu użyliśmy dokładnie tych samych atrybutów jak autorzy wyżej wymienionej pracy.

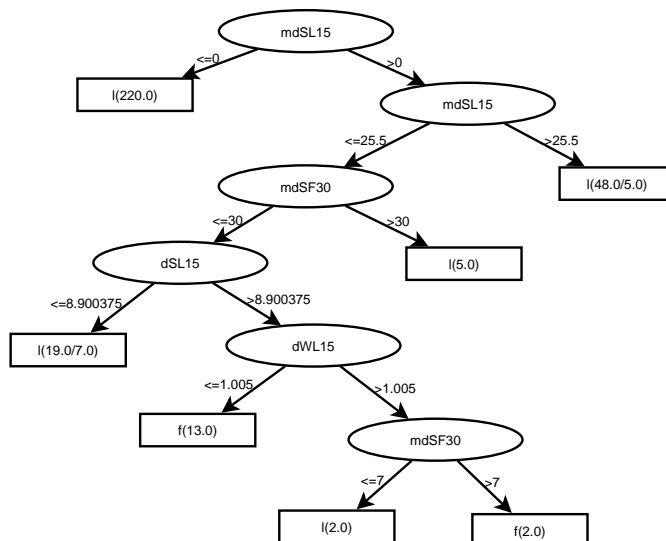
Atrybuty te to: **mdSFn**, **mdSLn** - mediana ceny przedmiotów sprzedanych w ciągu pierwszych/ ostatnich *n* dni aktywności, **mdWFn**, **mdWLn** - mediana ceny przedmiotów kupionych w ciągu pierwszych/ostatnich *n* dni aktywności, **dSFn**, **dSLn** - standardowe odchylenie ceny przedmiotów sprzedanych w ciągu pierwszych/ ostatnich *n* dni aktywności, **dWFn**, **dWLn** - standardowe odchylenie ceny

<sup>16</sup> <http://www.biskupi.pl>

Tabela 1. Macierz klasyfikacji

aktualne,przewidziane	legalny	oszust
oszust	48 %	52 %
legalny	4,2 %	95,8 %

przedmiotów kupionych w ciągu pierwszych/ostatnich  $n$  dni aktywności, **starte-dRatio** - stosunek ilości aukcji na których użytkownik był sprzedającym do wszystkich transakcji użytkownika. Uzyskane drzewo decyzyjne jest przedstawione na rys. 3.



Rysunek 3. Drzewo decyzyjne C4.5 dla podstawowych atrybutów

Proces walidacji poprawności modelu polegał na podziale zbioru uczącego na dwa podzbiory, każdy z nich raz używany był jako zbiór uczący, a następnie jako testowy. Macierz poprawności klasyfikacji przedstawiona jest w tabeli 1.

Atrybuty jakie zostały wybrane do stworzenia drzewa to: **mdSL15**, **mdSF30**, **dSL15**. Atrybut określający medianę ceny wystawionych na sprzedaż przedmiotów w ostatnich 15 dniach (**mdSL15**) ma największy wkład w klasyfikację. W naszym eksperymencie większe znaczenie miały atrybuty z krótszego przedziału czasowego (15 dni).

W naszym przypadku poprawność klasyfikacji oszustów nie jest imponująca i dużo niższa niż w eksperymencie amerykańskim (Chau (2005)), gdzie wynosiła około 80 %. Wygenerowana reguła na danych testowych dla oszustów sprawdza się jedynie w 48 %, co nie jest zadowalającym wynikiem. Gorsza skuteczność algorytmu może wynikać z tego, że amerykańskie realia aukcji internetowych różnią się od realiów polskich. Wśród różnic można wymienić między innymi fakt, że



rynek amerykański jest o wiele większy, co zapewnia oszustom większe pole do działania a także większą anonimowość. Ponadto, do rejestracji na amerykańskiej stronie serwisu eBay potrzeba numeru karty kredytowej, a na Allegro.pl procedura opiera się na dostarczeniu przez Allegro listu z kodem aktywacyjnym wysłanego na podany przez użytkownika adres. Czas potrzebny na rejestrację ma wpływ na to kiedy użytkownik zaczyna sprzedaż. Kolejną różnicą jest to, że klienci w obu krajach preferują inne formy płatności np. w Stanach Zjednoczonych najbardziej popularna jest płatność kartą kredytową oraz za pośrednictwem stron takich jak PayPal<sup>17</sup>, podczas gdy w Polsce ze względu na niską dostępność kart kredytowych ciągle popularne jest przesłanie przedmiotu za zaliczeniem lub wpłata na pocztę. Sposób rozliczenia ma wpływ na czas jaki zabiera sfinalizowanie transakcji. Biorąc pod uwagę powyższe obserwacje postanowiliśmy spróbować przeanalizować profil na polskich aukcjach i na tej podstawie wytypować atrybuty, które dawałyby lepsze wyniki podczas klasyfikacji.

### 5.3. Profile przestępców

Dzięki uzyskaniu danych o 27 kontaktach, z których zostały popełnione przestępstwa udało nam się przeanalizować zachowanie oszustów i opisać mechanizm przestępstw polegających na niedostarczeniu towaru na polskich aukcjach internetowych.

**Oszuści działają na aukcjach przez krótki okres.** Średni czas jaki upłynął od momentu zakończenia pierwszej aukcji, w której wzięli udział jako kupujący lub sprzedający do momentu zakończenia ostatniej wynosi średnio 51.6 dnia. Jeżeli chodzi o aukcje, gdzie byli sprzedającymi średni okres wynosi 45.5 dnia, a dla aukcji na których kupowali 13.9 dnia.

**W ostatnim okresie działalności oszustów można zaobserwować gwałtowny wzrost sprzedaży.** Niezależnie od tego czy oszust buduje reputację poprzez kupno czy sprzedaż, w ostatniej fazie jego działalności można zaobserwować znaczny wzrost liczby wystawionych przedmiotów.

Analizując zachowanie oszustów można stwierdzić, że:

- Oszuści w celu otrzymania komentarzy kupują przedmioty o małej wartości;
- Przedmioty sprzedawane przez oszustów zarówno w celu budowy reputacji jak i oszustwa mają podobną wartość i należą do podobnej kategorii.

### 5.4. Poszukiwanie nowych parametrów

Obserwacje zachowania oszustów, analiza ich zachowania jak i nie zadowalające wyniki wcześniejszych testów skłoniły nas do wyznaczenia bardziej adekwatnych atrybutów. Wprowadziliśmy dodatkowo kilka atrybutów opisujących oczekiwany zysk użytkownika oraz przyrost tego zysku w kolejnych okresach<sup>18</sup> działalności. Po- przez oczekiwany zysk rozumiemy kwotę jaką użytkownik uzyskałby, gdyby sprzedał wszystkie przedmioty wystawione przez niego na aukcjach w danym okresie

<sup>17</sup> <http://www.paypal.com>

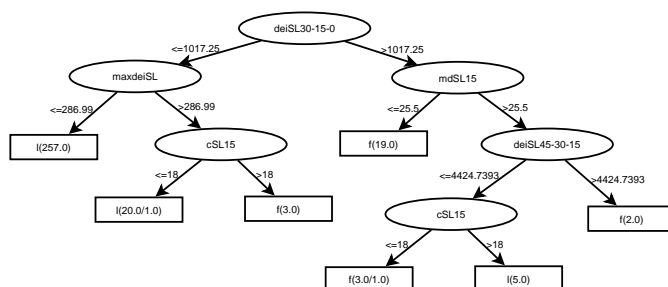
<sup>18</sup> Jako okres, w którym wyliczamy oczekiwany zysk, przyjęliśmy 15 dni

Tabela 2. Macierz klasyfikacji

aktualne, przewidziane	oszust	legalny
oszust	85 %	15 %
legalny	2,5 %	97,5 %

czasu. Atrybuty te mają odwzorować zauważony przez nas gwałtowny wzrost sprzedaży w ostatnich dniach działalności oszusta.

Te dodatkowe atrybuty to: **eiSL<sub>n</sub>**, **eiSF<sub>n</sub>** – zysk użytkownika z aukcji wystawionych w ostatnich/ pierwszych *n* dniach, **eiSL45-30** – zysk użytkownika z 15 dni od 45 do 30 dnia przed datą ostatniej aukcji, **eiSL30-15** – zysk użytkownika z 15 dni od 30 do 15 dnia przed datą ostatniej aukcji, **eiSL15-0** – zysk użytkownika w ostatnich 15 dniach, **deiSL45-30-15** – wzrost oczekiwanego zysku w przedostatnim okresie (różnica parametrów eiSL45-30 i eiSL30-15), **deiSL30-15-0** – wzrost oczekiwanego zysku w ostatnim (różnica parametrów eiSL30-15 i eiSL15-0), **maxdeiSL** – maksymalna wzrost oczekiwanego zysku (maksymalna wartość z parametrów deiSL45-30-15 i deiSL30-15-0). W kolejnym przeprowadzonym eksperymencie użyto wyżej przedstawionych parametrów i parametrów z eksperymentu amerykańskiego z pominięciem okresów 30 dni. Dodatkowymi wytypowanymi parametrami były: średnia cena oraz ilość sprzedawanych/kupowanych przedmiotów dla pierwszych i ostatnich 15 dni, oznaczone na drzewach odpowiednio literami **m** i **c**. Używając wyżej wymienionych parametrów algorytm C4.5 zbudował drzewo widoczne na rys. 4.



Rysunek 4. Drzewo decyzyjne C4.5 dla nowych atrybutów

Proces walidacji poprawności modelu przebiegał tak samo jak w poprzednim przypadku. Trafność klasyfikacji wyniosła 96,5%. Macierz poprawnego klasyfikowania przedstawiona jest w tabeli 2.

Tym razem klasyfikacja oszustów i legalnych użytkowników znacznie się poprawiła. Szczególnie widzimy to w przypadku oszustów, obecnie trafność klasyfikacji oszustów wynosi 85%, gdy poprzednio wynosiła zaledwie 48%. Możemy zauważyć, iż w drugim eksperymencie nowo wprowadzone atrybuty zostały uwzględnione przez algorytm i użyte do budowy reguł, co więcej znajdują się w korzeniu drzewa<sup>19</sup>.

<sup>19</sup> szczególnie **deiSL45-30-15** i **maxdeiSL**

Jedną z zasad budowania drzewa w oparciu o algorytm C4.5 jest wybór atrybutów na zasadzie minimalizacji entropii (zwiększania ilości informacji), czyli w korzeniu drzewa (poddrzewa) znajduje się zawsze atrybut z największym wpływem na klasyfikację. Nowo wprowadzone atrybuty okazały się zatem bardzo przydatne. W trakcie pierwszych testów na nowych zbiorach danych, algorytm wskazał kilku podejrzanych użytkowników z których dwóch po głębszej analizie okazało się być prawdziwymi przestępcami. Użytkownicy ci pasowali do odkrytego przez nas wzorca, np. w ostatniej fazie działalności chcieli zarobić jak najwięcej. Nie znaleźli się oni w naszym zbiorze uczącym, gdyż poszukiwaliśmy profili oszustów głównie poprzez ilość negatywnych komentarzy. Posiadali ich oni zaledwie kilka, gdyż prawdopodobnie ich oszustwo zostało udaremnione w początkowej fazie. Obaj użytkownicy mają w tej chwili zablokowane konta. To typowanie świadczy o tym iż profil oszustów został dobrze określony przez drzewo i może być z powodzeniem stosowany do wyznaczania podejrzanych użytkowników.

## 6. Podsumowanie i plany dalszego rozwoju

W ramach pracy opisanej w powyższym artykule udało nam się stworzyć system, potrafiący pobrać dane o aukcjach i użytkownikach z heterogenicznych źródeł. Na podstawie danych potrafi zbudować model klasyfikujący użytkowników aukcji, w zależności od ich dotychczasowej działalności. Zbudowany model można wykorzystać do analizy nowo ściągniętych danych. Otrzymane wyniki klasyfikacji można uznać za satysfakcjonujące. Udało nam się poprawić skuteczność algorytmu klasyfikującego. Wskazane przez nas parametry okazały się być trafne i zostały uwzględnione w klasyfikacji. Za sukces można też uznać zbudowanie rozszerzalnej architektury umożliwiającej prowadzenie dalszych badań.

W ramach dalszych prac nad systemem przewidywane jest umożliwienie wspierania innych portali aukcyjnych. Większość portali działających w Polsce działa na podobnych zasadach tj. stosuje system komentarzy, oraz analogiczny mechanizm przebiegu aukcji. Dzięki zastosowaniu ontologii udało się oddzielić dziedzinę problemu od konkretnego systemu aukcyjnego. Zadanie rozszerzenia sprowadza się więc do dobudowania mechanizmów pobierania danych z innych portali. Wartościowe byłoby również dalsze poszukiwanie optymalnych atrybutów, które mogłyby być podstawą klasyfikacji użytkowników. Można przeanalizować także inne algorytmy klasyfikacyjne. Wykorzystując dane przechowywane w ontologii można analizować również, powiązania pomiędzy użytkownikami aukcji pod kontem występowania podejrzanych wzorców.

## Literatura

- CHAU D. H., PANDIT S., and FALOUTSOS C. (2005) *Fraud detection in electronic auction..* In European Web Mining Forum at ECML/PKDD.
- CHAU D. H., PANDIT S., and FALOUTSOS C. (2005A) *Detecting fraudulent personalities in networks of online auctioneers.* In Proc. ECML/PKDD.

- CHAU D. H, PANDIT S., and FALOUTSOS C. (2007) *NetProbe: A Fast and Scalable System for Fraud Detection in Online Auction Networks*.
- CHAE M, SHIM S, CHO H, and LEE B. (2007) *An Empirical Analysis of Fraud Detection in Online Auctions: Credit Card Phantom Transaction*. HICSS.
- WANG J H, and CHIU C Q. (2005) *Detecting online auction inflated-reputation behaviors using Social Network Analysis*.
- GEMIUS SA (2006) *Internet 2006*  
[http://pliki.gemius.pl/Raporty/2006/Gemius\\_SA\\_Internet\\_2006.pdf](http://pliki.gemius.pl/Raporty/2006/Gemius_SA_Internet_2006.pdf)
- MONEY.PL *Internetowy biznes skazany na sukces*.  
<http://manager.money.pl/strategie>
- SMG/KRC *NetTrack*
- STEINER D (2005) *Interview with an eBay Vigilante*  
<http://auctionbytes.com/cab/abu/y205/m02/abu0136/s02>

## Analysis of Unlawful Behaviour Patterns in Internet Auctions

Internet transactions constitute an increasingly large part of the total turnover. The most important part of Internet commerce now is based on electronic auctions. Unfortunately, the popularity of auctions attracts fraudsters who prey on honest users. This paper concerns the problem of detecting and predicting fraudulent behaviour of Internet auction participants. The authors give an overview of the kinds of crimes committed using auction services and previous publications concerning their discovery. Then, the approach developed is presented. It embraces mechanisms for collecting data about auctions stemming from different sources and methods of storing them using an ontological representation. The methods previously used to detect fraudulent behaviour in Internet auctions were estimated under special conditions characteristic of Polish internet auctions. This analysis lead to a proposition of new coefficients to increase the quality of results given by an algorithm based on the decision trees method. The thesis is supported by a series of experimental tests.