Identyfikacja użytkowników sieci

przegląd metod i problemów

Krystian Baniak

07.10.2008

Agenda

- Wprowadzenie
- Identyfikacja – dziedzina zastosowania
- Rodzaje identyfikacji
- Techniki identyfikacji użytkowników sieci
- Problemy i wyzwania
- Dyskusja

Identyfikacja użytkowników sieci

Wprowadzenie

Użytkownik :: unikalna osoba, która ma dostęp do Internetu lub do innej sieci IP

Identyfikator :: zestaw informacji unikalnie związany z podmiotem; użytkownikiem

Profil :: zestaw cech charakterystycznych dla użytkownika

Identyfikacja użytkowników sieci

Dlaczego jest potrzebna?

- Identyfikacja ma na celu ustalenie dla danego typu aktywności powiązania z danym podmiotem w postaci obiektu:: identyfikatora określającego tożsamość podmiotu bądź profil jego działania.
- Zastosowanie technik identyfikacji
  - Uwierzytelnienie / potwierdzenie tożsamości
  - Personalizacja odpowiedzi systemu z którego korzysta podmiot
  - Rozliczalność transakcji wykonywanych w systemie przez podmiot
  - Przeprowadzanie analiz aktywności podmiotu w systemie
- identyfikacja w sensie udziału podmiotu
  - Aktywna, np. Proces logowania do systemu (z udziałem podmiotu)
  - Pasywna, np. system nadzoru obserwując aktywność i przebieg wydarzeń dokonuje identyfikacji podmiotu (bez udziału podmiotu)

Identyfikacja użytkowników sieci

Dziedziny zastosowań

Identyfikacja pasywna poprzez obserwację ma zastosowanie w wielu aspektach zarządzania sieciami IP.

- Zarządzanie sieciami (network management)
  - Określanie trendów i przewidywanie obciążenia sieci
- Bezpieczeństwo sieci
  - Analiza behawioralna – Intrusion prevention
  - Analiza środowiska nad którym nie sprawujemy kontroli administracyjnej
- Profilowanie i klasyfikacja użytkowników
  - Na potrzeby akcji marketingowych
  - Optymalny dobór reklam

Identyfikacja użytkowników sieci

Rodzaje zastosowań w kontekście badań autora

- Identyfikacja poprzez pasywną obserwację aktywności sieciowej
- Pasywna analiza aktywności sieciowej implikuje pewne ograniczenia
  - Niepełny zestaw informacji
  - Analizujemy środowisko nad którym nie mamy kontroli
  - Brak informacji o zależności pomiędzy obserwowanymi zmiennymi losowymi
  - Obserwowana maszyna używana jest przez wielu użytkowników
- Kiedy stosujemy takie rozwiązania
  - System z którego korzystamy nie dokonuje aktywnej identyfikacji podmiotu (np. witryna vortalu internetowego, system IDS)
  - Wymogi prywatności i anonimowości nie pozwalają na interakcję lub inspekcję identyfikatorów lub danych personalnych
  - Systemy analizy bezpieczeństwa sieci lub aplikacji, które dokonują profilowania subskrybentów w celu lepszego poznania charakteru obserwowanego systemu i osiągnięcia podstawowego poziomu rozliczalności w sytuacji wystąpienia anomalii z winy podmiotu

Identyfikacja użytkowników sieci

Elementy składowe identyfikatorów

- Identyfikacja poprzez obserwację typowych cech charakterystycznych dla obserwowanego protokołu komunikacyjnego lub aplikacji
- Poziom sieci – identyfikator adresowy
  - Poziom warstwy fizyczny (MAC address) – wąskie zastosowanie (bridging domain)
  - Poziom warstwy sieci (IP address) – podstawowy identyfikator użytkownika sieci
    - Adresacja dynamiczna DHCP, adres migruje po nodach – DSL, DOCSIS
    - Translacja adresów N:M przekreśla zasadność stosowania tego identyfikatora
    - GPRS/EDGE/UMTS IP adres migruje pomiędzy MSISDN po destrukcji kontekstu PDP
    - Zazwyczaj składnik hybrydowego identyfikatora
  - Identyfikator hybrydowy
    - DSL/CABLE: stateful map IP-MAC (DHCP snooping, RADIUS accounting) (znaczenie lokalne)
    - GSM/UMTS: GGSN accounting based map MSISDN←>IP (MSISDN jest unikalny ale wymaga ingerencji w sieć operatora)

Podsumowując adres IP + podejście hybrydowe ma limitowane zastosowanie i wymaga spełnienia określonych kryteriów aby być wiarygodnym. W wielu zastosowaniach jest jednak wystarczające.

Identyfikacja użytkowników sieci

Przykłady

- System kontroli rodzicielskiej u operatora fixed lub mobilnego bazujący na inspekcji ruchu RADIUS [RADWARE, BLUECOAT]
  - Wymaga dostępu do specyficznej bazy danych i zaufania do systemu kontroli rodzicielskiej
- System kontroli dostępu do treści (multimedia) który wykorzystuje IP adres w celu ograniczenia odbiorców do obszaru geograficznego lub puli zarządzanej przez danego zarządcę [BBC ONLINE TV]
- System IPS/IDS który analizuje profil ruchu pochodzący z danego IP adresu i dostosowujący zabezpieczenia do profilu aktywności
  - DoS protection
  - SYN flood
  - Kwarantanna dla IP adresów przekraczających ograniczenia
  - Identyfikacja o niskiej jakości nie nastawiona na rozliczalność w sensie prawnym

Identyfikacja użytkowników sieci

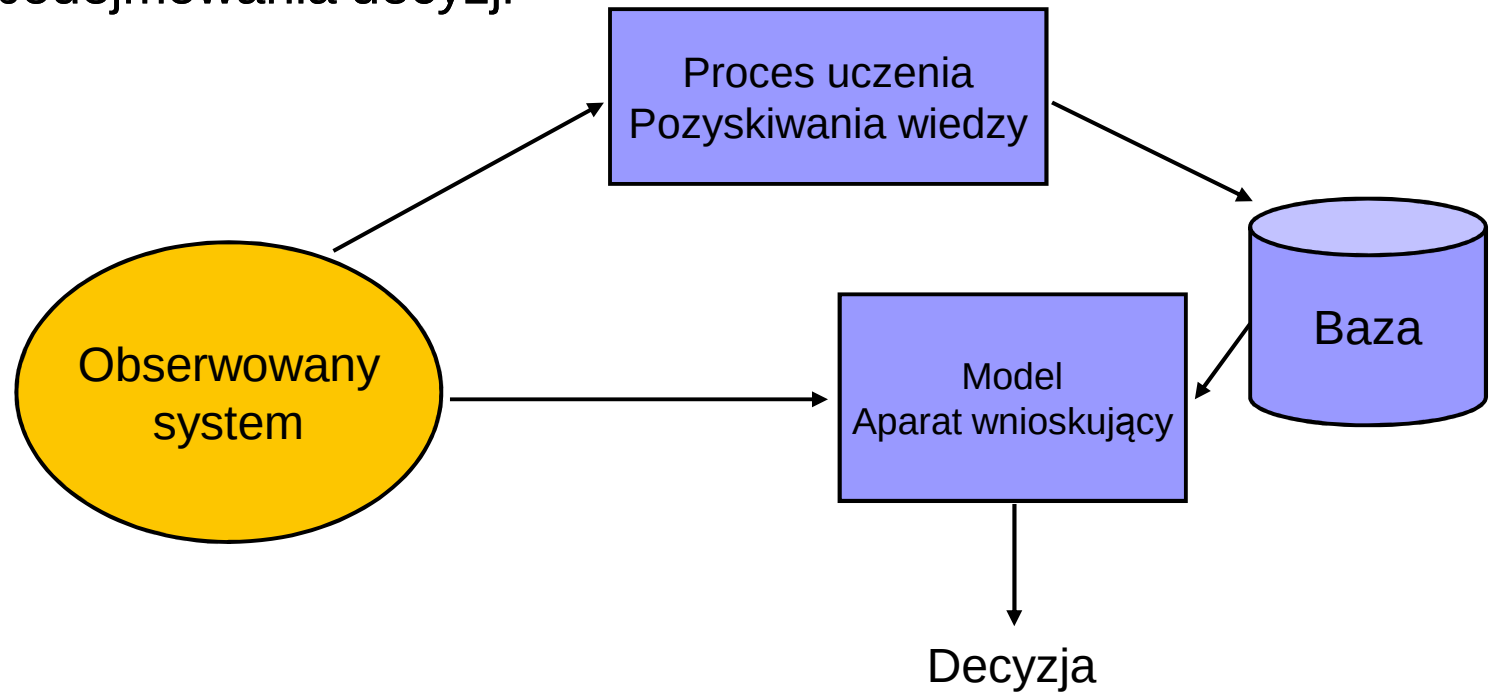
Elementy składowe identyfikatorów

- Poziom aplikacyjny i poziom sesji
  - Transakcje Web (40-50% ruchu Internetowego in total)
  - W powiązaniu z warstwą sieciową bardzo skuteczny
  - Analiza nagłówka HTTP
    - User-Agent
    - Referer
    - Cookie
    - URL
  - Klasyfikacja obiektu do którego odwołuje się zapytanie HTTP GET
    - GEOIP
    - URL rating (Sport|p0rn|news ….)
  - Wykorzystanie popularnych serwisów Internetowych (Google)
    - Wyniki zapytań – słowa kluczowe
- Elementy dodatkowe
  - Rozkład aktywności w czasie

Identyfikacja użytkowników sieci

techniki pozyskiwania wiedzy

- Aby zaimplementować skuteczny system identyfikacji należy zadbać o odpowiednie metody pozyskiwania wiedzy, którą wykorzystamy do podejmowania decyzji



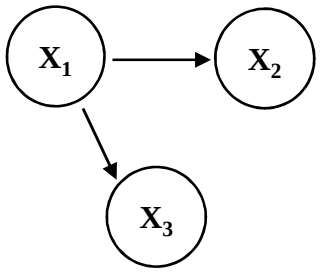
Identyfikacja użytkowników sieci

Reprezentacja wiedzy – wybrane metody

- Sieci Bayesa
  - Doskonale nadają się do modelowania w stanie niepewności
  - Modelujemy zależności pomiędzy zmiennymi losowymi, które są elementami naszego profilu
  - Bazując na dowodach i znanemu prawdopodobieństwu ich występowania możemy poznać prawdopodobieństwo wystąpienia danego wydarzenia w powiązaniu z obserwowanymi dowodami w przyszłości

$$p(A \mid B) = \frac{p(A, B)}{p(B)} = \frac{p(B \mid A)p(A)}{p(B)}$$

$$p(A_i \mid E) = \frac{p(E \mid A_i)p(A_i)}{p(E)} = \frac{p(E \mid A_i)p(A_i)}{\sum_i p(E \mid A_i)p(A_i)}$$



$p(x_1, x_2, x_3) = p(x_3 \mid x_1)\, p(x_2 \mid x_1)\, p(x_1)$

Identyfikacja użytkowników sieci

Reprezentacja wiedzy – wybrane metody

- Sieci Bayesa problemy
  - Złożoność obliczeniowa rosnąca wykładniczo wraz z ilością parametrów wejściowych (obserwowanych zdarzeń losowych)
  - Proces uczenia skomplikowany, wymaga stosowania heurystyk

- Przykłady

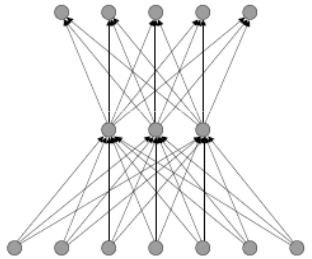
S.K.M. Wong C.J. Butz „A Bayesian Approach to User Profiling in Information Retrieval”

- Dane wejściowe: zbiór dokumentów oraz użytkowników
- Wykorzystują sieć Bayesa do tworzenia profilu preferencji użytkownika co do wyboru i przydatności dokumentów
- Zastosowanie np. filtrowanie wiadomości pocztowych w skrzynce odbiorczej
- P(Relevance=true| A1=a1, A2=a2, …)

Identyfikacja użytkowników sieci

Reprezentacja wiedzy – wybrane metody

- Sieci Neuronowe
  - Parametry obserwowanego obiektu są wejściem dla sieci neuronowej
  - Kluczowym jest dobór odpowiedniego zestawu uczącego
  - Idealne dla modelowania zjawisk opisywanych przez dziesiątki parametrów



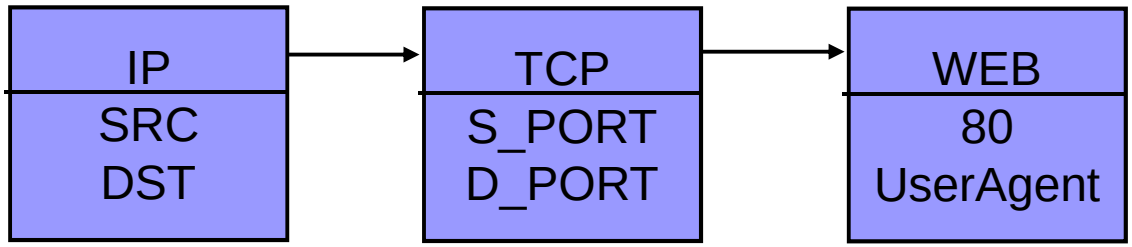
Przykład:

- **Tom Auld, Andrew W. Moore „**Bayesian Neural Networks For Internet Traffic Classification”
- Sieć neuronowa służąca do identyfikacji typu(p2p,mail,web,…) obserwowanego przepływu sieciowego (TCP)
- Skuteczność ~99% (email) i ~95% po 8 miesiącach
- Sieć neuronowa, 246 znormalizowanych(<0,1>) parametrów obserwowanych statystyczne cechu sesji TCP
- Cecha charakterystyczna – duże zestawy danych treningowych, uczenie techniką Bayesa

Identyfikacja użytkowników sieci

Reprezentacja wiedzy – wybrane metody

- Ramy (ang. Frames, Marvin Minsky, MIT)
  - Zapis wiedzy polegający na stereotypowaniu zjawisk; wydarzeń
  - Ramy wykorzystują uporządkowanie hierarchiczne
  - Rama opisuje sytuację, która ma uwarunkowanie (wartości parametrów) oraz przewiduje następstwa wynikające z
  - Nie modelują czasu ani zmian modelu w czasie
  - Umożliwiają formalizację zapisu co w połączeniu z odpowiednią notacją umożliwia wykorzystanie wyników w dalszej analizie



Identyfikacja użytkowników sieci

Reprezentacja wiedzy – wybrane metody

- Logika rozmyta (fuzzy logic)
  - Logika wielowartościowa
  - Modelowanie niepewności, niejednoznaczności
  - Często stosowana z alg. ewolucyjnymi i sieciami neuronowymi
- Przykład
  - Radware DefensePro
  - Analiza behawioralna i tworzenie modelu sieci
  - Model decyzyjny oparty na logice rozmytej
- metody hybrydowe
  - Stosowane do modelowania zjawisk o dużej komplikacji
  - Różne metody używane do normalizacji i formalizacji różnych parametrów

Identyfikacja użytkowników sieci

Problemy i wyzwania

- obserwowanie użytkowników sieciowych korzystających z NAT
- obserwowanie użytkowników sieciowych korzystających z tej samej maszyny (np. komputer w bibliotece)
  - Charakterystyka ruchu sieciowego
  - Ruch agregowany ~Poisson(k,L)
- szyfrowanie i anonymizacja

Identyfikacja użytkowników sieci

Propozycja metody hybrydowej

- Rozproszony system agentowy profilujący użytkowników sieciowych
- Założenia
  - Rozliczalność na poziomie maszyny w sieci (node)
  - Wykrywanie maszyn współdzielonych
  - Wstępna Identyfikacja podmiotów na maszynach współdzielonych
- Wykorzystywane techniki
  - Ramki wykorzystywane do stereotypowania zachowań sieciowych
  - Zbudowanie ontologii aktywności sieciowej w celu uproszczenia wnioskowania o zachowaniu podmiotu obserwowanego
  - Identyfikacja w oparciu o metodę hybrydową
    - Parametry statystyczne aktywności sieciowej
    - Profil preferencji i zainteresowań
    - Zbiór unikalnych identyfikatorów aplikacyjnych

Identyfikacja użytkowników sieci

Przykład profilu preferencji podmiotu

```
:dom:
    Computers/Internet: 31
    nocat: 317
    Web Advertisements: 3
    ads: 25
    cleaning updatesites: 2
    Search Engines/Portals: 16
    Health: 1
  :nat:
    IT: 58
    "NO": 1
    FR: 10
    DE: 7
    HK: 5
    EU: 1
    CH: 1
    GB: 554
    IE: 112
    JP: 1
    PL: 495
    SE: 5
    CA: 1
    NL: 19
    US: 120
  :ua:
    Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1): 3
    Client: 1
    Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1): 389
    Windows-Update-Agent: 1
    contype: 1
total_stats:
  :timed:
    :slotD:
    - 0
    - 0
    :slotE:
    - 0
    - 0
    :slotA:
    - 0
    - 0
    :slotB:
    - 815
    - 8431643
    :slotC:
    - 180
    - 2298239
  :apps:
    :HTTPS: 115
    :HTTP: 131
    :SMTP: 749
  :hits: 995
  :size: 10729882
```

Identyfikacja użytkowników sieci

Dziękuję za uwagę