STUDIA OECONOMICA POSNANIENSIA 2013, vol. 1, no. 2(251)

Jakub Dzikowski
Uniwersytet Ekonomiczny w Poznaniu, Wydział Informatyki i Gospodarki Elektronicznej, Katedra Informatyki Ekonomicznej
jakub.dzikowski@kie.ue.poznan.pl

WYSZUKIWANIE DANYCH OSOBOWYCH W INTERNECIE DLA CELÓW INFORMATYKI ŚLEDCZEJ

Streszczenie: Popełnienie przestępstwa z wykorzystaniem infrastruktury Internetu jest łatwiejsze i obarczone mniejszym ryzykiem wykrycia niż popełnienie przestępstwa w tradycyjny sposób. Zwalczanie przestępczości w cyberprzestrzeni możliwe jest dzięki śladom (np. danym personalnym), jakie każdy użytkownik sieci zostawia. Artykuł omawia zagadnienie przestępczości w cyberprzestrzeni oraz zagadnienia związane ze zwalczaniem takiej przestępczości. Zaprezentowana została metoda wyszukiwania danych osobowych w Internecie, obejmująca automatyczne rozszerzenie posiadanych informacji o poszukiwanym, a także empiryczna ocena skuteczności tej metody.

Słowa kluczowe: model użytkownika, informatyka śledcza, ślad użytkownika, *Page Rank*.

Klasyfikacja JEL: C88, D83.

Wstęp

Wraz z rozwojem środowiska biznesowego i społecznego pojawiają się nowe możliwości popełniania przestępstw [Wall 2005]. Przestępczość w cyberprzestrzeni informacyjnej wykorzystuje różne modele biznesowe [Rush i in. 2008], cechuje się mniejszym ryzykiem wykrycia w porównaniu z tradycyjną przestępczością, a w dodatku umożliwia osiągnięcie większych zysków [Cárdenas i in. 2009]. Obecne ataki w cyberprzestrzeni nie są przeprowadzane już tylko dla zabawy przez hobbystów, ale w coraz większym stopniu przez przestępców i szpiegów [Hypponen 2007]. O ile dawniej głównymi motywami przestępstw powiązanych z technologiami informacyjnymi były: ciekawość, edukacja oraz poszukiwanie sławy, o tyle teraz do cyberprzestępstw prowadzą: chęć finansowego zysku, zemsta lub pobudki polityczne [Choo, Smith i McCusker 2007]. W znacznym stopniu z przestępczością w sieci są związane także zorganizowane grupy przestępcze [Choo i Smith 2008].

Przestępcy działający w Internecie kierują się złudnym wyobrażeniem o anonimowości, tymczasem każda aktywność internetowa zostawia za sobą ślad (ang. *footprint*). Aby do tego śladu dotrzeć, można przesłuchać poszukiwanego lub osoby

z nim związane, przeprowadzić analizę śledczą jego komputera lub przeszukać Internet [Garfinkel i Cox 2009]. W przypadku ostatniego z tych sposobów wystarczy wpisać nazwę lub pseudonim poszukiwanego jako zapytanie do standardowej wyszukiwarki (Google, Yahoo!, Bing) lub wyszukiwarki na portalu społecznościowym. Wykorzystując techniki przetwarzania informacji i danych, można zautomatyzować proces pozyskiwania informacji ze śladów pozostawionych przez daną osobę w sieci [Abramowicz i in. 2010; Dzikowski 2010].

W artykule zaprezentowano metodę pozyskiwania informacji ze śladów użytkownika sieci, która, wykorzystując techniki ekstrakcji i wyszukiwania informacji, pozwala na utworzenie reprezentacji poszukiwanego zawierającej pożądane informacje. W pierwszej sekcji przedstawiono rodzaje przestępczości związanej z Internetem oraz sposoby jej zwalczania. Następnie opisano informatykę śledczą, dyscyplinę z pogranicza informatyki oraz prawa, formalizującą sposoby obchodzenia się z dowodem elektronicznym. W kolejnych sekcjach przedstawiono techniki istotne z punktu widzenia tworzonej metody: wyszukiwanie oraz ekstrakcję informacji, a także modelowanie użytkownika. Następnie została zaprezentowana metoda pozyskiwania informacji ze śladów internetowych wraz z przeprowadzonymi badaniami trafności pozyskiwanych informacji. Artykuł kończy się wnioskami z przeprowadzonych badań oraz propozycjami ulepszeń utworzonej metody.

1. Przestępczość związana z Internetem

Przestępczość w sieci jest różnorodna. Niektóre z przestępstw są ekstrapolacją przestępstw rzeczywistych, inne to zjawiska całkiem nowe. Różne są sceny, obszary cyberprzestrzeni, w których zjawiska te mają miejsce, na różnych poziomach też mogą one być zwalczane. Kolejne podsekcje posłużą usystematyzowaniu zjawisk związanych z przestępczością w sieci.

1.1. Rodzaje przestępczości w sieci

Można wyodrębnić trzy rodzaje przestępczości powiązanej z technologiami informacyjnymi [Wall 2005]: dotyczącą integralności, przeciwko mieniu oraz dotyczącą zawartości. Poniżej opisano każdy z tych rodzajów, uwzględniając rolę Internetu.

Przestępczość dotycząca integralności
Przestępczość dotycząca integralności to najmniej szkodliwy rodzaj przestępczości powiązanej z Internetem [Wall 2005], aczkolwiek często stanowi wstęp do przestępczości bardziej szkodliwej i bardziej złożonej [Rush i in. 2008; Wall 2005]. Niektóre z przestępstw istniejących w sieci (na przykład łamanie zabezpieczeń sieci telekomunikacyjnych, takie jak *phreaking* [Rajagopalan 2000]) dzięki Internetowi zyskały większe możliwości, jednak gdyby go nie było, przestępstwa te istniałyby

dalej – tak jak istniały w epoce przedinternetowej i istnieją obecnie – z wykorzystaniem infrastruktury innych systemów telekomunikacyjnych [Wall 2005]. Nowe możliwości dla tradycyjnej przestępczości niosą ze sobą łamanie zabezpieczeń systemów informatycznych (*hacking* oraz *cracking*), wirusy komputerowe lub haktywizm (czyli wykorzystanie *hackingu* w celach ideologicznych).

Nowe typy przestępstw, możliwe dzięki Internetowi, a dotyczące integralności komputera, to wysyłanie spamu, wojna informacyjna (realizująca się na przykład przez utrudnienie dostępu do informacji lub propagandę) czy też ataki mające na celu zakłócenie działania systemów informatycznych (Denial of Service lub Parasitic Computing).

Komputerowe przestępstwa przeciwko mieniu

Internet ułatwia popełnianie oszustw finansowych [Canhoto 2007], w szczególności zastosowanie schematów piramidy finansowej i sprzedaży lawinowej [Pareja 2008]. Całkiem nowe możliwości zyskały istniejące przestępstwa przeciwko mieniu: pojawiły się nowe rodzaje oszustw, na przykład oszustwa nigeryjskie[1] i kradzież tożsamości[2], łatwiejsze stało się popełnianie kradzieży tajemnicy handlowej [Wall 2005]. Spośród całkiem nowych rodzajów przestępstw należy wymienić kradzież własności intelektualnej (piractwo komputerowe), hazard w sieci (w państwach, w których jest on zabroniony) albo oszustwa na serwisach aukcyjnych [Wall 2005]. Dzięki wykorzystaniu infrastruktury Internetu oraz możliwości automatyzacji niektórych działań, możliwe są zarówno oszustwa drobne, jak i oszustwa na masową skalę [Cárdenas i in. 2009].

Przestępczość dotycząca zawartości

Przestępstwa dotyczące zawartości mogą być podzielone na dwie grupy. W pierwszej znajdują się przestępstwa związane z obscenicznością [Gray 2009]. Druga grupa jest związana z treściami, które są wymierzone w konkretną osobę (jako pewna forma przemocy) [Wall 2005].

Wśród przestępstw związanych z obscenicznością należy przede wszystkim wymienić pewne rodzaje obrotu materiałami erotycznymi i treściami pornograficznymi, które w ramach konkretnej jurysdykcji mogą być niedozwolone. Przestępstwami w większości państw są: uwodzenie nieletnich, rozprzestrzenianie pornografii dziecięcej, cyberprostytucja itp. [Chatterjee 2005]. W drugiej grupie przestępstw dotyczących zawartości znajdują się na przykład nawoływanie do nienawiści, prześladowanie lub stalking[3], propagowanie faszyzmu, zorganizowane rozmowy o produkcji substancji niedozwolonych (bomb, narkotyków).

[1] Oszustwo nigeryjskie (ang. *419 scam*) to takie oszustwo, w ramach którego ofiara zostaje wciągnięta w fikcyjny transfer znacznej kwoty pieniędzy [Dyrud 2005].

[2] Czyli podszycie się pod kogoś, celowe wykorzystanie danych osobowych innej osoby [Bilge i in. 2009].

[3] Stalking – złośliwe i powtarzające się nagabywanie, naprzykrzanie, nękanie [Deirmenjian 1999].

Tabela 1. Macierz przestępczości w cyberprzestrzeni według D.S. Walla

	Przestępstwa dotyczące integralności (szkodliwe wykroczenia)	**Komputerowe przestępstwa przeciwko mieniu**	**Przestepstwa dotyczące zawartości (obsceniczność)**	**Przestępstwa dotyczące zawartości (przemoc)**
Większe możliwości dla tradycyjnej przestępczości	oszustwa telekomunikacyjne (*phreaking*)	oszustwa; piramidy finansowe (sprzedaż lawinowa)	obrót materiałami o charakterze erotycznym i pornograficznym	stalking; prześladowanie
Nowe możliwości dla tradycyjnej przestępczości	cracking, haking; wirusy komputerowe; haktywizm	złożone oszustwa na wielką skalę; oszustwa nigeryjskie (*419 scams*); kradzież tajemnicy handlowej; kradzież tożsamości	czerpanie korzyści z wykonywania czynności seksualnych przed kamerami internetowymi (strony z *camgirls*, *camboys*)	nawoływanie do nienawiści; zorganizowana pedofilia (wykorzystywanie nieletnich)
Nowe możliwości dla nowych rodzajów przestępstw	spam; ataki Denial of Service; wojna informacyjna; *Parasitic Computing*	kradzież własności intelektualnej; hazard w sieci; oszustwa na serwisach aukcyjnych; drobne oszustwa na masową skalę	cyberseks; sutenerstwo w Internecie	grooming; wymiana informacji dotyczących produkcji bomb, narkotyków; nawoływanie do nienawiści skierowane do konkretnych osób

Źródło: Wall [2005].

Jako podsumowanie, w tabeli 1 przedstawiono macierz przestępczości w cyberprzestrzeni: konkretne przykłady przestępstw, z uwzględnieniem rodzaju przestępstwa oraz roli, jaką w ich pojawieniu się odegrał Internet.

1.2. Sceny przestępstw w Internecie

Przestępstwa mogą dotyczyć różnych obszarów cyberprzestrzeni. W szczególności należy wyróżnić sieciowe gry komputerowe, portale społecznościowe, inne serwisy, w których treść jest tworzona przez użytkowników, oraz portale aukcyjne i ogłoszeniowe. W obszarach tych może dochodzić do popełnienia przestępstwa lub też przestępstwa mogą się w nich manifestować.

Sieciowe gry komputerowe
W masowych sieciowych grach komputerowych (ang. *Massively Multiplayer Online Role-Playing Game*, MMORPG) tworzone są wirtualne światy, w których gracz prezentowany jest poprzez awatar [Cole i Griffiths 2007]. Gracz ma możliwość wymiany waluty świata rzeczywistego na wirtualne pieniądze, obowiązujące w grze. Tym samym większość oszustw finansowych Internetu i świata rzeczywistego może być przeniesiona w niektóre gry MMORPG [Choo i Smith 2008].

Ponieważ gry tego typu umożliwiają interakcję pomiędzy graczami i do pewnego stopnia odtwarzają świat rzeczywisty, możliwe są także inne przestępstwa, na przykład sprzedaż nieistniejących artefaktów innym graczom, napaść i kradzież cennych przedmiotów, funkcjonowanie kręgów pedofilskich [Whitty, Young i Goodings 2011]. Inne, bardziej klasyczne przestępstwa to inżynieria zwrotna (ang. *reverse engineering*) i modyfikacja kodu gry w celu zwiększenia zysków albo kradzież tożsamości (przejęcie awatara) [Choo i Smith 2008].

Portale społecznościowe
Portal społecznościowy (ang. *social networking site*) to usługa wykorzystująca sieć, która umożliwia użytkownikom tworzenie publicznych lub częściowo publicznych profili, określanie relacji z innymi użytkownikami oraz przeglądanie istniejących relacji [Boyd i Ellison 2007]. W popularnych portalach społecznościowych, jak Facebook, Nasza-Klasa, LinkedIn, użytkownicy mogą zamieszczać informacje dotyczące ich prywatności, podawać dane personalne, zamieszczać osobiste zdjęcia. Informacje te są przydatne dla przestępców [Choo i Smith 2008] i mogą być wykorzystane w kradzieży tożsamości, stalkingu, a także w innych szkodliwych celach [Govani i Pashley 2007]. Portale społecznościowe mogą być także wykorzystane w celach propagandowych lub do rozprowadzania złośliwego kodu. Często także zamieszczane tam treści naruszają prawa autorskie [Choo i Smith 2008].

Chcąc chronić obywateli, niektóre jednostki policji w Stanach Zjednoczonych tworzą własne konta na portalach społecznościowych, zakładając, że jeśli ktoś ma wśród znajomych komendę policji, jest mniej narażony na czyhające na portalach społecznościowych zagrożenia [Miller 2010]. Oprócz tego przeprowadzane są szkolenia, jak wykorzystać serwisy społecznościowe w podczas prowadzenia śledztwa.

Treści tworzone przez użytkowników
W epoce Web 2.0 to użytkownicy tworzą treść sieci. Tym samym mogą zamieszczać na przykład treści pornograficzne i pedofilskie, porady, jak zbudować bombę, jak popełnić samobójstwo, materiały wideo zawierające przemoc i innego rodzaju informacje, które nie powinny być prezentowane bez kontroli lub wcale nie powinny być prezentowane [George i Scerri 2007]. Często także treści zamieszczane przez użytkowników w rzeczywistości nie są ich autorstwa. W serwisie YouTube znaczna część klipów wideo to zmodyfikowany obraz z podkładem muzycznym, gdzie naruszane są prawa autorskie zarówno do obrazu, jak i do muzyki. Jest to jednak robione

na tak wielką skalę, że użytkownicy mogą nawet być nieświadomi nielegalności tego typu procederu [George i Scerri 2007].

Dodatkowo, w Internecie mogą być zamieszczane treści naruszające prywatność innych osób lub stanowiące zagrożenie dla ich własnej prywatności [George i Scerri 2007] albo treści propagandowe, nawołujące do nienawiści [Choo i Smith 2008]. Łatwo naruszyć czyjąś reputację, zamieszczając odpowiedni post na blogu lub modyfikując artykuł w Wikipedii, podając mniej lub bardziej prawdziwe informacje [George i Scerri 2007].

Portale aukcyjne i ogłoszeniowe
Serwisy aukcyjne mogą być wykorzystywane przez zorganizowane grupy przestępcze na przykład w celu prania brudnych pieniędzy [Choo i Smith 2008], poprzez dokonywanie transakcji po zawyżonych cenach. Trywialnymi przypadkami jest zakładanie aukcji z produktami o zaniżonych cenach, a następnie niedostarczanie towaru, albo też zakładanie fikcyjnych sklepów internetowych [Antes i in. 2008]. Ponadto przedmiot obrotu może być niedopuszczony do sprzedaży, naruszać prawa innych osób fizycznych lub prawnych, może stanowić niebezpieczeństwo dla klientów lub być oferowany przez sprzedawcę nieposiadającego wymaganych uprawnień.

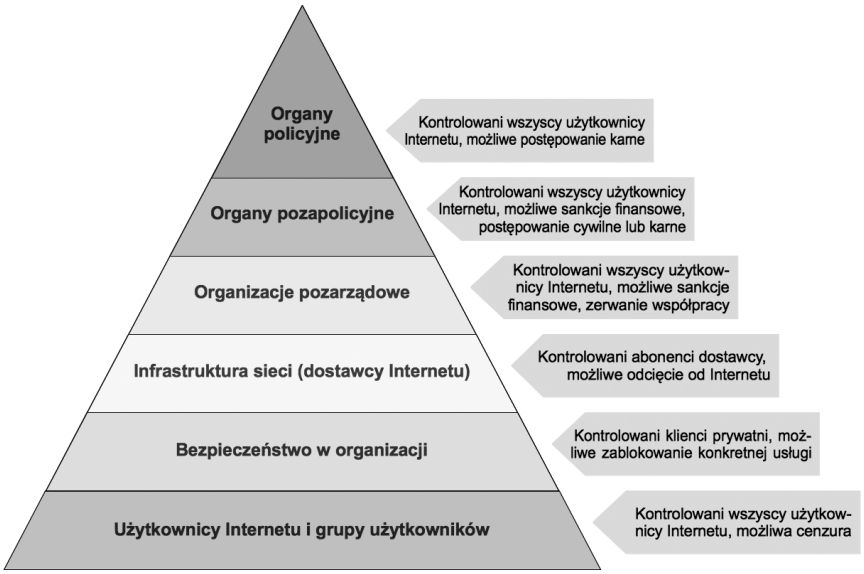
Podobne zagadnienia dotyczą portali z ogłoszeniami, które dodatkowo nie ponoszą żadnej odpowiedzialności za poprawne dokonanie transakcji i nie mają protokołów weryfikacji tych transakcji.

1.3. Zwalczanie przestępczości związanej z Internetem

Popularność usług wykorzystujących Internet drugiej generacji (Web 2.0), nacisk na współpracę użytkowników przez sieć, tworzenie treści przez użytkowników i wsparcie dla internetowych społeczności – wszystko to skutkuje nowymi sposobami wymiany informacji oraz nowymi wyzwaniami dla systemów prawnych [Choo i Smith 2008; George i Scerri 2007]. Problemem jest odkrycie tożsamości podejrzanego, ponieważ osoba popełniająca czyn niedozwolony najczęściej występuje pod pseudonimem. Ze względu na to, że sieć jest globalna, a prawo ma zasięg lokalny, często trudno stwierdzić, którą jurysdykcją należy się kierować w postępowaniu i kwalifikacji czynu. W przypadku niektórych treści trudno określić, czy coś jest już nielegalne, czy jeszcze takie nie jest. Wreszcie niektóre ograniczenia mają naturę praktyczną – kontrola cyberprzestrzeni jest trudna, łatwo popełnić przestępstwo z jej wykorzystaniem, a usunięcie niedozwolonych treści rzadko jest wystarczające w jej zwalczaniu.

W sieci działają różne grupy osób oraz organizacje, które są odpowiedzialne za kontrolę działań użytkowników Internetu oraz publikowanych treści [Wall 2005]. Użytkownicy Internetu i grupy użytkowników (na przykład moderatorzy na forach internetowych) dokonują cenzury, a w razie konieczności mogą powiadomić odpowiednie inne podmioty. W ramach zapewnienia bezpieczeństwa w organizacji

może zostać podjęte wiele działań, aby narzucić określone praktyki, włączając w to blokady konkretnych dostarczanych usług. Dostawcy Internetu mogą zaprzestać świadczenia usług telekomunikacyjnych. Niektóre organizacje pozarządowe mogą nałożyć sankcje finansowe lub zerwać współpracę. Organy pozapolicyjne mogą wykorzystywać kombinację kar, grzywien oraz gróźb postawienia w stan oskarżenia. Organy policyjne, czyli oprócz policji na przykład Europol albo Interpol, chociaż odgrywają najmniejszą rolę (jeśli chodzi o liczbę przypadków), mają do dyspozycji najbardziej dotkliwe sankcje.



Rysunek 1. Piramida kontroli cyberprzestrzeni

Źródło: Dzikowski [2010]

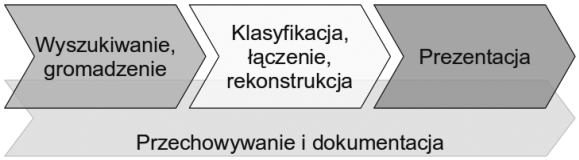
Opisane powyżej poziomy kontroli oraz możliwe sankcje można przedstawić na piramidzie kontroli cyberprzestrzeni (rysunek 1). Niższym poziomom kontroli odpowiada większa liczba przypadków mniej poważnych, z kolei wyższe poziomy kontroli dysponują bardziej dotkliwymi sankcjami. W przypadku gdy na niższym z poziomów nie uda się rozwiązać określonego problemu związanego z przestępczością w cyberprzestrzeni (lub niepożądanymi treściami), a problem jest wystarczająco poważny, może zostać delegowany do wyższej warstwy. Niezależnie od tego, podmioty znajdujące się w konkretnych warstwach mogą przekazywać sobie informacje o przejawach przestępczości lub domniemanych przestępcach, a także postulować nałożenie sankcji. Na przykład agencje rządowe w ramach zwalczania przestępczości w cyberprzestrzeni zgłaszają korporacji Google, aby zablokowane zostały niektóre treści, albo zwracają się o udostępnienie danych dotyczących określonego użytkownika usług Google. W Brazylii i Indiach

wnioski agencji rządowych dotyczą usuwania treści w popularnym, będącym własnością Google, serwisie społecznościowym Orkut; od niemieckich agencji Google otrzymuje listy adresów stron internetowych, które powinny zostać zablokowane w Niemczech, ze względu na naruszane tam prawo [*Google. Raport przejrzystości 2012*].

2. Informatyka śledcza

Informatyka śledcza (ang. *computer forensics*, *digital forensics*, *forensic computing*) jest dziedziną wiedzy na pograniczu nauk sądowych oraz informatycznych. Obejmuje gromadzenie oraz analizę danych z systemów informatycznych, sieci, komunikacji bezprzewodowej oraz nośników danych w taki sposób, aby zgromadzone dane mogły być uznane w sądzie za dowód [US-CERT 2008]. Kluczowe zatem jest pojęcie dowodu elektronicznego (cyfrowego, ang. *digital evidence*), który można opisać jako przechowywana lub przesyłana informacja w postaci elektronicznej, która ma znaczenie dowodowe [Królikowski 2009]. Większość dowodów elektronicznych, przez swą ulotność, wrażliwość na upływ czasu, łatwość uszkodzenia, skopiowania lub spreparowania, najczęściej może mieć jedynie charakter poszlaki. Informatyka śledcza opisuje takie sposoby postępowania z dowodem elektronicznym, aby nie naruszyć jego integralności [McKemmish 2008].

Proces śledczy w informatyce śledczej można przedstawić jako następujące po sobie etapy wyszukiwania oraz gromadzenia dowodów cyfrowych, ich analizy (klasyfikacji, łączenia i rekonstrukcji) oraz prezentacji (rysunek 2) [Pollitt 2008; Solomon, Broom i Barrett 2004; US-CERT 2008]. Równolegle należy dokumentować wszelkie operacje dokonane na dowodzie elektronicznym.



Rysunek 2. Proces śledczy w informatyce śledczej

Źródło: Dzikowski [2010]

Niektórzy autorzy sugerują, że najlepiej gromadzić wszystko, co może być nawet pośrednio związane z przypadkiem [Solomon, Broom i Barrett 2004], inni podkreślają ogrom danych, które powinny być przeanalizowane, i że obecnie używane narzędzia informatyki śledczej nie są w stanie sobie poradzić z analizą wszystkiego [Ayers 2009]. Klasyfikacja, łączenie oraz rekonstrukcja są w artykule ujmowane zbiorczo jako analiza dowodu elektronicznego. Zgodnie z postulatami informatyki śledczej, analiza ta powinna być wykonywana jedynie na kopii danych z nośnika,

przy czym zgodność tych danych z danymi oryginalnymi powinna być sprawdzona przed analizą (np. poprzez porównanie sum kontrolnych obu kopii) [Solomon, Broom i Barrett 2004]. Analiza może być dokonywana na różnych poziomach abstrakcji dowodu elektronicznego [Carrier 2003]. Na przykład warstwami abstrakcji dla dokumentu HTML są: sektor dysku, partycja, system plików, plik, ciąg znaków ASCII, kod HTML (tabela 2). Większość obecnie stosowanych narzędzi informatyki śledczej ciągle jeszcze skupia się na poziomie systemu operacyjnego, a nie na poziomach wyższych, jak poziom aplikacji [Altheide i Carvey 2011; Ayers 2009; Casey 2011; Liu i in. 2012].

Tabela 2. Warstwy abstrakcji dowodu elektronicznego

Warstwa	Przedmiot analizy
Dane fizyczne	analiza nośników danych, odzyskiwanie danych
Zarządzanie danymi	przechowywanie danych, podział na partycje, organizacja dysków (warstwa nie musi zawsze występować)
System plików	analiza zawartości plików
Aplikacje	badanie plików z uwzględnieniem ich wykorzystania w aplikacjach; analiza dzienników zdarzeń, plików konfiguracyjnych, obrazów, dokumentów
Sieć	niskopoziomowa analiza pakietów przesyłanych przez sieć (analiza dzienników zdarzeń usług sieciowych lub serwerów to analiza na poziomie aplikacji)
Pamięć	analiza niezapisywanych na stałe, wrażliwych danych, wykorzystywanych przez programy

Źródło: Carrier [2003].

Podstawowym celem, jeśli chodzi o przechowywanie dowodów, jest możliwość zapewnienia, że żadne zmiany nie zostały dokonane na dowodach od momentu ich pozyskania [Solomon, Broom i Barrett 2004]. W świecie cyfrowym polega to na izolacji potencjalnych dowodów i zgromadzeniu informacji, które mogą zostać utracone podczas lub po zakończeniu pracy systemu [Carrier 2003]. Dokumentować należy wszystkie działania, które były wykonane na dowodzie elektronicznym, żeby w razie konieczności cały proces można było odtworzyć. Tym samym wszystkie akcje związane identyfikacją, gromadzeniem, zabezpieczaniem, transportem i przechowywaniem powinny zostać udokumentowane [Mukasey, Sedgwick i Hagy 2008].

Celem prezentacji dowodów jest przekonanie odbiorców (sędziego, prokuratora, grupy menedżerów) do wykorzystania dowodów, czyli tym samym przekonanie, że prezentowany dowód potwierdza jeden lub więcej faktów [Solomon, Broom i Barrett 2004].

3. Wyszukiwanie i ekstrakcja informacji

Wyszukiwanie informacji (ang. *information retrieval*, IR) często jest przeprowadzane na dużej kolekcji nieustrukturyzowanych dokumentów, a jego wyniki mają zaspokoić potrzebę informacyjną [Manning, Raghavan i Schütze 2008]. Formalną reprezentacją potrzeby informacyjnej są kryteria wyszukiwania, które w wypadku wyszukiwarek internetowych podawane są w odpowiednim polu formularza na stronie WWW.

W dokumentach elektronicznych odnalezionych z wykorzystaniem systemów wyszukiwawczych można poszukiwać określonych informacji. Służy do tego ekstrakcja informacji (ang. *information extraction*) definiowana jako dowolny proces, w którym dochodzi do selektywnej strukturyzacji i łączenia danych [Cowie i Lehnert 1996], wskutek którego dane są interpretowane i stają się informacją [Abramowicz 2008]. Prostym przykładem ekstrakcji informacji może być wyszukanie w dokumencie adresu e-mail. Kolejnym, bardziej zaawansowanym przykładem może być pozyskanie z tego dokumentu informacji, do kogo ten adres należy.

Jak już wspomniano, informatyka śledcza ciągle jeszcze zazwyczaj skupia się na niższych warstwach abstrakcji dowodów elektronicznych, o ekstrakcji można natomiast mówić dopiero na poziomie warstwy aplikacji, gdy aplikacje automatycznie analizują nagłówki dokumentów albo dzienniki zdarzeń. Większość z wymienionych funkcjonalności mają Forensic Toolkit (FTK) oraz EnCase (czyli najpopularniejsze komercyjne narzędzia informatyki śledczej), jednak trudno mówić o automatyzacji – nawet z wykorzystaniem tak zaawansowanych platform procesy wyszukiwania i ekstrakcji informacji w znacznej mierze wykonywane są przez człowieka[4]. Tymczasem ekstrakcja informacji w systemach niezwiązanych z informatyką śledczą i przestępczością w cyberprzestrzeni jest zazwyczaj znacznie bardziej zaawansowana.

4. Użytkownik w sieci

Informacje o użytkowniku sieci mogą być podane bezpośrednio przez niego, przez osoby trzecie, albo też mogą być zgromadzone automatycznie podczas interakcji z określonym systemem. Tworzone są reprezentacje użytkownika, zwane modelami lub profilami użytkownika, które zawierają informacje dotyczące jego cech i kontekstu.

4.1. Model użytkownika i modelowanie użytkownika

W. Wahlster i A. Kobsa [1989] definiują model użytkownika jako źródło wiedzy[5] w systemie zawierającym sprecyzowane przypuszczenia dotyczące wszystkich

[4] Por. funkcjonalności wspomnianych narzędzi: http://accessdata.com/products/computer-forensics/ftk oraz http://www.guidancesoftware.com/forensic.htm [dostęp: 18.04.2012].

[5] Wiedza jest tutaj rozumiana jako wiedza w systemach reprezentacji wiedzy i systemach sztucznej inteligencji; w kontekście systemów informatycznych należałoby posługiwać się raczej pojęciem informacji.

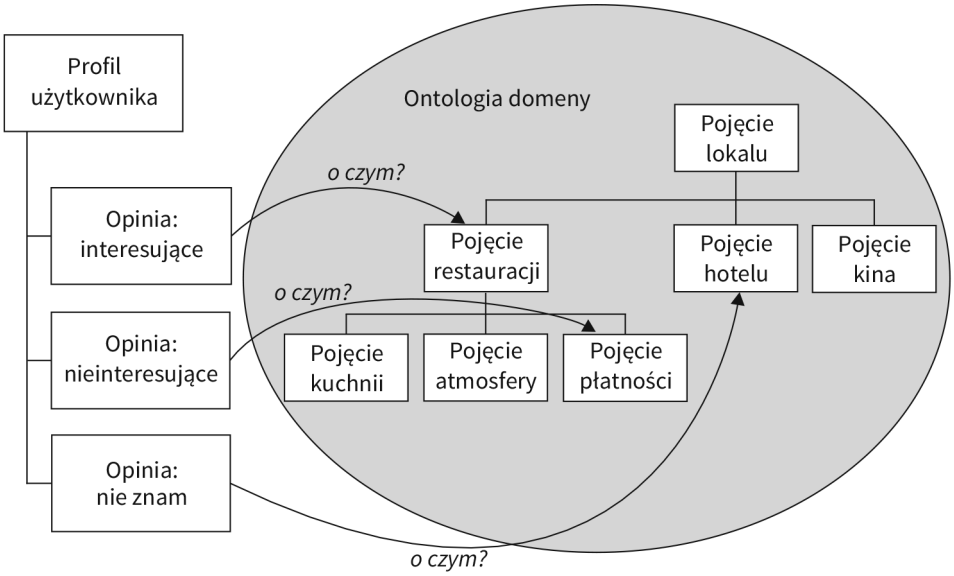
aspektów użytkownika, które mogą być przydatne z punktu widzenia działania systemu. W przypadku nieposiadania informacji o użytkowniku, system zachowywałby się identycznie dla każdego użytkownika [Fröschl 2005]. Model użytkownika jest zatem niezbędny wszędzie tam, gdzie oczekuje się zindywidualizowanej odpowiedzi systemu.

Można wyróżnić metody implicite oraz explicite gromadzenia informacji o użytkowniku. Metody implicite nie angażują bezpośrednio użytkownika (informacje o nim pozyskiwane są automatycznie), natomiast metody explicite wymagają jego ingerencji (na przykład poprzez wypełnienie odpowiedniego pola w formularzu) [Abramowicz 2008].

4.2. Wewnętrzna struktura modelu

Jako najczęstszą reprezentację modelu wskazuje się kolekcję słów kluczowych [Gauch i in. 2007]. Mogą one być pozyskane automatycznie (na przykład na podstawie zawartości przeglądanych stron WWW) lub podane wprost przez użytkownika. Ponadto słowa kluczowe można opatrywać wagami określającymi poziom zainteresowania tematem.

Modele użytkownika mogą być również reprezentowane przez sieć semantyczną (ang. *semantic net*, *semantic network*). Służy ona do reprezentacji wiedzy oraz wnioskowania [Sowa 2006]. Sieć semantyczną można przedstawić w postaci grafu, którego wierzchołki reprezentują pojęcia, a krawędzie – powiązania pomiędzy tymi



Rysunek 3. Model użytkownika typu „overlay"

Źródło: Gawinecki [2005]

pojęciami. Dodatkowo, krawędzie oraz wierzchołki mogą być opatrzone wagami ilustrującymi ich ważność [Gauch i in. 2007]. Modele użytkownika zbudowane z semantycznych sieci mają przewagę nad modelami zbudowanymi ze słów kluczowych, ponieważ mogą jednoznacznie określać relacje pomiędzy określonymi słowami lub pojęciami [Gauch i in. 2007].

Model może być również opisany przez ontologię. Definiuje się ją jako ogólne słowa oraz pojęcia użyte do opisu i reprezentacji dziedziny wiedzy [Daconta, Obrst i Smith 2003; Guarino 1998]. Charakterystycznymi rodzajami pojęć składającymi się na ontologię są klasy (byty generalne), instancje (byty szczegółowe), relacje pomiędzy tymi bytami, ich właściwości (atrybuty), dotyczące tych bytów funkcje, procesy, ograniczenia oraz reguły [Daconta, Obrst i Smith 2003]. Ontologia może w kompleksowy sposób opisywać pewną dziedzinę wiedzy, także wiedzę o użytkowniku posiadaną przez system.

W modelu typu „overlay" wiedza użytkownika reprezentowana jest jako podzbiór wiedzy systemu, przy czym wiedza systemu to wiedza ekspercka lub dziedzinowa [Koch 2000], która może być zamodelowana na przykład w postaci słów kluczowych, sieci semantycznej lub ontologii (rysunek 3). Brak wiedzy eksperckiej można wywnioskować na podstawie porównania wiedzy użytkownika z wiedzą systemu [Henze i Nejdl 2003].

4.3. Publicznie dostępne informacje o użytkowniku

W portalach społecznościowych użytkownicy tworzą swoje profile, udostępniając publicznie przynajmniej część informacji o sobie [Boyd i Ellison 2007]. Podobnie udostępniają informacje, prowadząc własną, osobistą stronę internetową (ang. *personal home page*). Na takiej stronie użytkownik opisuje, kim jest, podaje swoje dane kontaktowe, zainteresowania itp. [Chandler 1998]. Dodatkowo informacje o sobie użytkownik sieci może zamieszczać w prowadzonych przez siebie blogach internetowych, na forach internetowych oraz w komentarzach do artykułów tworzonych przez innych autorów. Dane te są raczej w niewielkim stopniu ustrukturyzowane.

Informacje o użytkowniku sieci mogą też być udostępniane przez inne osoby. Jako przykład należy podać informacje o pracowniku przedsiębiorstwa, którego imię, nazwisko, służbowy adres e-mail i telefon podane są na stronie internetowej przedsiębiorstwa. Innym przykładem może być podanie danych pewnej osoby na własnym blogu lub we własnym wpisie na forum. Ze względu na popularność określonych osób, ich dane osobowe mogą być podawane także na innych stronach internetowych.

Istotnym miejscem, w którym mogą się znajdować informacje o użytkowniku, są zasoby „głębokiego Internetu" [Kaczmarek 2006], czyli zasoby, które nie są indeksowane przez standardowe wyszukiwarki, a dostęp do nich możliwy jest zazwyczaj jedynie z wykorzystaniem formularzy. Ze względu na sposób prezentacji, na przykład automatyczne generowanie stron internetowych z wykorzystaniem rekordów z bazy danych, informacje te są zazwyczaj dobrze ustrukturyzowane.

5. Proponowane rozwiązanie

Rozwiązanie prezentowane w artykule wykorzystuje techniki wyszukiwania i ekstrakcji informacji w procesie pozyskiwania danych poszukiwanej osoby dla celów informatyki śledczej. Rozwiązanie to wpisuje się w etapy wyszukiwania i gromadzenia oraz klasyfikacji, łączenia i rekonstrukcji. Najpierw wyszukiwane są dokumenty z danymi osobowymi poszukiwanej osoby, a następnie budowany jest jej model.

Podczas opracowywania rozwiązania konieczne okazały się następujące działania:

- wybór struktury modelu poszukiwanej osoby,
- ustalenie źródeł informacji o poszukiwanym i sposobu wyszukania wartości, które mogą być danymi osobowymi,
- opracowanie sposobu szacowania wiarygodności odnalezionych wartości,
- opracowanie sposobu wybierania wartości, które ostatecznie powinny się znaleźć w modelu i służyć jego rozbudowaniu.

W tej sekcji opisano proponowane rozwiązanie pod kątem przedstawionych wyzwań.

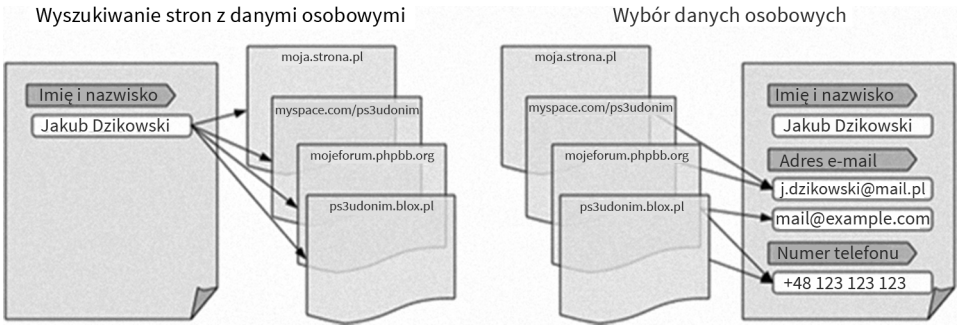
5.1. Struktura modelu

Model poszukiwanego jest rodzajem modelu typu „overlay". Model dziedziny stanowi kolekcja wszystkich odnalezionych dokumentów i znajdujących się w nich wartości, które mogą być danymi osobowymi poszukiwanego. Są one połączone w sieć semantyczną z jednokierunkowymi relacjami, której pojęcia są opatrzone rangami. Rangi reprezentują wiarygodność odnalezionych dokumentów i wartości. Dodatkowo wszystkie wartości odnalezione w dokumentach przyporządkowane są do przynajmniej jednej z kategorii, na przykład „Imię i nazwisko", „Adres e-mail", „Nr telefonu".

Warstwą wiedzy użytkownika w modelu są te dane osobowe, które zostały uznane za wystarczająco wiarygodne. Klasa odnalezionych wartości w warstwie wiedzy użytkownika, które należą do pewnej kategorii, jest nazywana atrybutem poszukiwanego. W modelu zdefiniowane są atrybuty poszukiwanego, których wartości mają być wyszukiwane, a każdy atrybut może być realizowany przez dowolną liczbę wartości atrybutów.

5.2. Wyszukiwanie danych osobowych

W prezentowanym rozwiązaniu wyszukiwane dane osobowe pochodzą z ustrukturyzowanych w niewielkim stopniu źródeł Internetu widocznego. Najpierw, z wykorzystaniem odpowiedniej wyszukiwarki, pozyskiwane są dokumenty zawierające wartości atrybutów poszukiwanego, na przykład jego imię i nazwisko (rysunek 4). Znane wartości atrybutów użytkownika traktowane są jako kryteria wyszukiwania.



Rysunek 4. Wyszukiwanie danych osobowych

Ponadto, w niektórych przypadkach podawane są także kryteria pochodne – na przykład dla imienia i nazwiska „Jakub Dzikowski" dodatkowym kryterium będzie także „J. Dzikowski".

Następnie z wykorzystaniem odpowiednich technik ekstrakcji informacji (na przykład wyrażeń regularnych) z odnalezionych dokumentów wyodrębniane są potencjalne wartości atrybutów poszukiwanego (np. ciągi znaków, które pasują do zdefiniowanego wzorca na adres e-mail). Ponieważ wiele z odnalezionych danych w rzeczywistości nie będzie wartościami atrybutów poszukiwanego, dodatkowo szacowana jest przez system wiarygodność tych danych, aby ustalić, które z nich powinny się znaleźć w modelu użytkownika.

Podczas wyszukiwania budowana jest sieć będąca modelem dziedziny w modelu poszukiwanego. Dodawane są do niej wartości, kryteria pochodne oraz odnalezione dokumenty, łączone z innymi pojęciami w sieci relacją jednokierunkową (z *A* do *B*) w następujących przypadkach:

– z wykorzystaniem wartości atrybutu poszukiwanego (*A*) odnaleziono dokument (*B*),
– z wartości atrybutu (*A*) wyprowadzono wartość pochodną (*B*), która później może być wykorzystana w dalszym wyszukiwaniu,
– w dokumencie (*A*) odnaleziono możliwą wartość atrybutu (*B*).

Wyszukiwanie wykonywane jest iteracyjnie, a w każdej iteracji wykorzystywane są wszystkie wartości atrybutów poszukiwanego. Tym samym odnalezione w poprzedniej iteracji wartości mogą zostać wykorzystane w rozbudowie modelu w iteracji bieżącej.

5.3. Ocena wiarygodności odszukanych danych osobowych

Większość zidentyfikowanych w dokumentach potencjalnych wartości atrybutów nie jest poprawnymi danymi poszukiwanego. Odpowiednie wzorce ekstrakcji wyodrębniają z dokumentów ciągi znaków, które mogą być danymi osobowymi, jednak nie rozstrzygają, czy dane te pasują do danych znajdujących się już w modelu. W celu

obliczenia rangi świadczącej o wiarygodności wartości atrybutów użytkownika posłużono się zmodyfikowaną wersją algorytmu PageRank, dawniej wykorzystywanego w wyszukiwarce Google [Brin i Page 1998]. Algorytm ten pozwala na nadanie rang połączonym w sieć pojęciom w zależności od relacji pomiędzy nimi. Ranga jest tym większa, im większa jest liczba relacji prowadzących do określonego pojęcia i im większe rangi mają pojęcia, z których wychodzą te relacje.

Niech $R(v,t)$ oznacza wartość rangi pojęcia v w iteracji t, $M(v)$ – zbiór pojęć, z których prowadzą relacje do pojęcia v, a $L(v)$ – liczbę relacji wychodzących z pojęcia v. Ponadto niech N oznacza liczbę wszystkich pojęć, a d tzw. współczynnik tłumienia, który decyduje o polaryzacji rang (gdy $d = 0$, wszystkie rangi przyjmują wartości $1/N$, natomiast gdy $d = 1$, rangi są najbardziej zróżnicowane). Wtedy, zgodnie z publicznie znaną wersją algorytmu PageRank [Brin i Page 1998], w kolejnych iteracjach rangi obliczane są zgodnie ze wzorem:

$$R_{t+1}(v_i) = \frac{1-d}{N} + d \sum_{v_j \in M(v_i)} \frac{R_t(v_j)}{L(v_j)}. \tag{1}$$

Ponieważ w prezentowanym rozwiązaniu korzystna jest różna polaryzacja rang (na przykład mniejsza dla nazw miejscowości, a większa dla adresu e-mail), została wprowadzona możliwość manualnego ustalania wartości parametru d dla różnych atrybutów użytkownika. Dodatkowo, niektóre wartości atrybutów, na przykład podane przez śledczego imię i nazwisko poszukiwanego, powinny być traktowane jako pewne, a tym samym mieć najwyższą rangę. Dlatego sztucznie podwyższana jest ich ranga poprzez utworzenie dodatkowych relacji w tworzonej sieci – do wartości, która jest pewna, prowadzą relacje wychodzące ze wszystkich pozostałych pojęć.

5.4. Wybór „właściwych" wartości

Aby wybrać te spośród odnalezionych wartości, które powinny się znaleźć w modelu, wykorzystano metodę naturalnego podziału, która pozwala na odnalezienie największych skoków w wartościach rang i wyznaczenie granicznej minimalnej rangi. Jeżeli zachodzą warunki: $v'_i \notin C(V)$, $v'_i \in A_k$, $R_i \geqslant 1/n$, $i = 1, 2, \ldots, n_k$ oraz $R_1 \geqslant R_2 \geqslant \ldots \geqslant R_{n_k}$, gdzie n jest liczbą wszystkich zasobów, $C(V)$ jest zbiorem zasobów zatwierdzonych przez śledczego, A_k zbiorem wartości atrybutu k, to dla j, takiego że $1 \leqslant j < n_k$, minimalna ranga dla wartości atrybutu k określana będzie wzorem:

$$R^k_{\min} = \arg\max_{R_j} \left(\frac{R_j}{R_{j+1}} \right). \tag{2}$$

Podczas przeprowadzanych eksperymentów okazało się, że metoda ta, stosowana zazwyczaj w analizie skupień do wyznaczenia liczby skupień, umiarkowanie trafnie oddzielała wartości atrybutów użytkownika od pozostałych wartości, które nie powinny się znaleźć w modelu.

6. Przeprowadzone badania i wynikające z nich wnioski

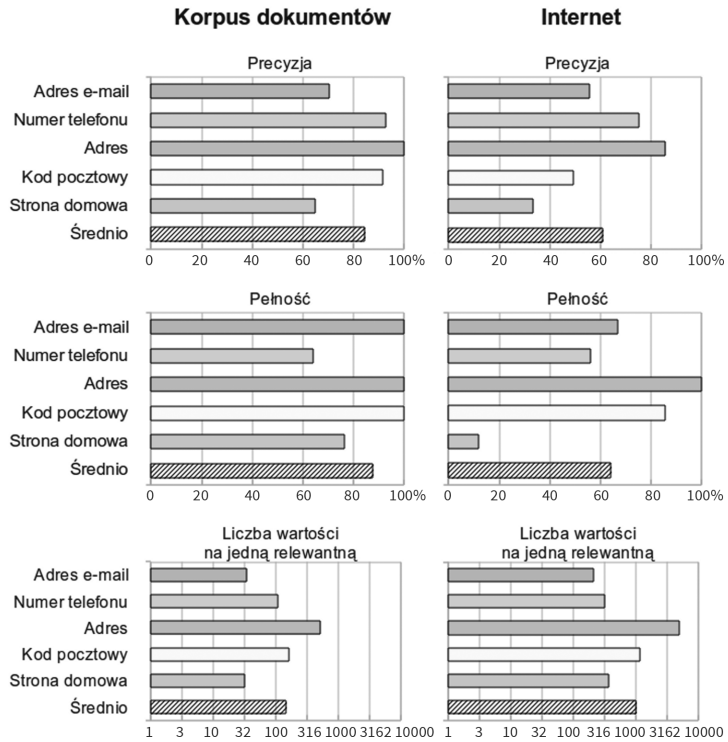
Aby ocenić skuteczność prezentowanego rozwiązania, przeprowadzono badania polegające na wyszukiwaniu danych osób, które prowadzą swoje strony domowe. Założeniem było, że zawsze znane jest imię i nazwisko poszukiwanej osoby. Poszukiwano natomiast adresów e-mail, numerów telefonów, adresów (ulicy i numeru), kodów pocztowych oraz adresów stron domowych wybranych osób. Odpowiednie wartości współczynników tłumienia ustalono w ten sposób, aby najbardziej spolaryzowane były rangi imienia i nazwiska, adresu e-mail, numeru telefonu oraz adresu. Najmniej zróżnicowane miały być rangi dla atrybutu „strona domowa".

Przygotowano korpus składający się z 1323 dokumentów wyszukanych dla fraz „adwokat Poznań", „notariusz Warszawa" itp. Wybrano wszystkie te imiona i nazwiska, które występowały przynajmniej w 10 dokumentach i dla których możliwe było manualne odnalezienie poszukiwanych danych. Następnie automatycznie budowano modele wybranych siedmiu poszukiwanych osób, wykorzystując jedynie dokumenty korpusu oraz zasoby sieci (poprzez wyszukiwarkę Google).

Badano relewancję odnalezionych wartości, wyrażoną przez pełność (stosunek liczby wartości relewantnych w modelu poszukiwanego do wszystkich wartości relewantnych) oraz precyzję (stosunek liczby wszystkich wartości w modelu do wszystkich relewantnych w modelu). Dodatkowo wprowadzoną miarą była liczba wszystkich odnalezionych wartości w danej kategorii (nie tylko wartości atrybutu) przypadająca na jedną wartość relewantną. Wyniki badań zaprezentowano na rysunku 5.

Dla wyszukiwania w korpusie uzyskano średnią precyzję na poziomie 82% oraz średnią pełność 87% przy 151 wartościach przypadających na jedną wartość relewantną. Najmniej trafne okazało się wyszukiwanie adresów stron domowych. W przypadku wyszukiwania w Internecie osiągnięte wyniki okazały się znacznie gorsze niemal dla wszystkich atrybutów poszukiwanego. Średnia precyzja wyniosła 61%, pełność 64%, a na jedną wartość relewantną przypadało średnio 770 wartości nierelewantnych. Spadła trafność wyszukiwania danych osobowych przy jednoczesnym zwiększeniu ilości danych podlegających analizie.

Przeprowadzone badania pozwoliły na zidentyfikowanie kilku problemów, z którymi rozwiązanie powinno się zmierzyć. Występowanie powtarzających się danych, na przykład takie samo nazwisko różnych osób, powoduje obniżenie trafności budowania modelu. Negatywnie na otrzymane wyniki wpływa także zbyt mała ilość informacji o poszukiwanym. Zdarzało się, że w określonej iteracji w modelu znajdowały się wszystkie pożądane informacje oraz jedna wartość atrybutu innej osoby. Ta wartość czasami prowadziła do tego, że w kolejnych iteracjach wyszukiwania pojawiało się coraz więcej wartości atrybutów drugiej osoby. Szczególnie złe wyniki dla wyszukiwania adresu strony domowej wynikają z trudności w przygotowaniu właściwego wzorca ekstrakcji dla tego atrybutu. W większości przypadków wraz ze wzrostem liczby wartości przypadających na jedną wartość relewantną następował spadek precyzji i pełności wyszukiwania.



Rysunek 5. Wyniki przeprowadzonego eksperymentu

Źródło: Dzikowski [2010]

Można postulować przede wszystkim przeprowadzenie bardziej szczegółowych badań, dla większej liczby osób i z wykorzystaniem innych niż Google wyszukiwarek internetowych, które być może pozwoliłyby dotrzeć do większej ilości informacji. Należałoby także przeprowadzić kompleksowe badania dla innych niż wykorzystane w tych badaniach wzorców ekstrakcji, takich, które mogą służyć do zidentyfikowania informacji trudnych do wyodrębnienia z tekstu.

Prezentowana metoda sprawdziła się umiarkowanie dobrze w budowaniu modelu zawierającego informacje o zainteresowaniach użytkownika [Abramowicz i in. 2010].

Podsumowanie

W artykule przedstawiono metodę automatycznego budowania modelu użytkownika dla potrzeb wspomagania zwalczania przestępczości w sieci. Przedstawiono rodzaje, sceny oraz metody zwalczania przestępczości w cyberprzestrzeni, a następnie przydatne z punktu widzenia techniki tworzonego rozwiązania. Dokonano weryfikacji

zaprezentowanego rozwiązania dla wyszukiwania danych osób mających swoje strony internetowe.

Dalsze prace powinny się skupić na udoskonaleniu wzorców ekstrakcji dla poszczególnych danych osobowych. Do rozwiązania powinny zostać włączone źródła wykraczające poza strony HTML Internetu widocznego. Korzystne mogłoby się okazać przeszukiwanie Internetu głębokiego, innych typów plików lub wykorzystanie wyszukiwarek, które mają dostęp do zastrzeżonych treści na portalach społecznościowych albo wewnętrznych baz danych organizacji.

Bibliografia

Abramowicz, W., 2008, *Filtrowanie informacji*, Wydawnictwo Akademii Ekonomicznej w Poznaniu, Poznań.

Abramowicz, W., Dzikowski, J., Filipowska, A., Małyszko, J., Węckowski, D.G., 2010, *Odkrywanie tożsamości użytkownika sieci*, Zastosowania Systemów Informatycznych Zarządzania, Uniwersytet Warszawski, Wydział Zarządzania.

Altheide, C., Carvey, H., 2011, *Digital Forensics with Open Source Tools*, Syngress, Waltham, USA.

Antes, J., Conley, J., Morris, R., Schossow, S., Yee, Z., Fang, F., 2008, *Cyber Crimes: Real Life and in the Virtual World*, http://public.csusm.edu/fangfang/Teaching/HTMmaterial/StudentProjectFall08/Groupl-paper.pdf [dostęp: 18.04.2012].

Ayers, D., 2009, *A Second Generation Computer Forensic Analysis System*, *Digital Investigation*, vol. 6, s. 34–42.

Bilge, L., Strufe, T., Balzarotti, D., Kirda, E., 2009, *All Your Contacts are Belong to Us: Automated Identity Theft Attacks on Social Networks*, w: *Proceedings of the 18th International Conference on World Wide Web*, ACM, 20–24 kwietnia 2009, Madrid, Spain, s. 551–560.

Boyd, D., Ellison, N., 2007, *Social Network Sites: Definition, History, and Scholarship*, Journal of Computer Mediated Communication, vol. 13, no. 1, s. 210.

Brin, S., Page, L., 1998, *The Anatomy of a Large-scale Hypertextual Web Search Engine*, Computer Networks and ISDN Systems, vol. 30, no. 1–7, s. 107–117.

Canhoto, A., 2007, *Profiling Behaviour: The Social Construction of Categories in the Detection of Financial Crime*, rozprawa doktorska, University of London.

Cárdenas, A., Radosavac, S., Grossklags, J., Chuang, J., Hoofnagle, C., 2009, *An Economic Map of Cybercrime*, w: *TPRC: The 37th Research Conference on Communication, Information and Internet Policy*, 26–28 April 2009, Arlington, VA, USA.

Carrier, B., 2003, *Defining Digital Forensic Examination and Analysis Tools Using Abstraction Layers*, International Journal of Digital Evidence, vol. 1, no. 4, s. 1–12.

Carrier, B., Spafford, E.H., 2003, *Getting Physical with the Digital Investigation Process*, Tech. Rep., Center for Education and Research in Information Assurance and Security, Purdue University.

Casey, E., 2011, *Digital Evidence and Computer Crime: Forensic Science, Computers, and the Internet*, Academic Press.

Chandler, D., 1998, *Personal Home Pages and the Construction of Identities on the Web*, http://leahanderst.com/Composition/Personal\%20Home\%20Pages\%20and\%20the\%20Construction\%20of\%20Identities\%20on\%20the\%20Web.pdf [dostęp: 18.04.2012].

Chatterjee, B., 2005, *Pixels, Pimps and Prostitutes: Human Rights and the Cyber-Sex Trade*, Routledge Cavendish, s. 11–26.

Choo, K.K.R., Smith, R.G., 2008, *Criminal Exploitation of Online Systems by Organised Crime Groups*, Asian Criminology, vol. 3, s. 37–59.

Choo, K.K.R., Smith, R.G., McCusker, R., 2007, *Future Directions in Technology-enabled Crime: 2007–2009*, Tech. Rep., Australian Institute of Criminology.

Cole, H., Griffiths, M., 2007, *Social Interactions in Massively Multiplayer Online Role-playing Gamers*, CyberPsychology & Behavior, vol. 10, no. 4, s. 575–583.

Cowie, J., Lehnert, W., 1996, *Information Extraction*, Communications of the ACM, vol. 39, no. 1, s. 80–91.

Daconta, M.C., Obrst, L.J., Smith, K.T., 2003, *The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management*, Wiley.

Deirmenjian, J., 1999, *Stalking in Cyberspace*, Journal of the American Academy of Psychiatry and the Law Online, vol. 27, no. 3, s. 407–413.

Dyrud, M., 2005, *"I Brought You a Good News": An Analysis of Nigerian 419 Letters*, w: *Proceedings of the 2005 Association for Business Communication Annual Convention*, 20–25 October 2005, Irvine, CA, USA.

Dzikowski, J., 2010, *Wykrywanie przestępczości z wykorzystaniem informacji ze źródeł internetowych*, praca magisterska, Uniwersytet Ekonomiczny w Poznaniu.

Fröschl, C., 2005, *User Modeling and User Profiling in Adaptive E-learning Systems*, praca magisterska, Graz University of Technology.

Garfinkel, S., Cox, D., 2009, *Finding and Archiving the Internet Footprint*, w: *First Digital Lives Research Conference: Personal Digital Archives for the 21st Century*, London, England.

Gauch, S., Speretta, M., Chandramouli, A., Micarelli, A., 2007, *User Profiles for Personalized Information Access*, Springer-Verlag, Berlin-Heidelberg, s. 54–89.

Gawinecki, M., 2005, *Modelowanie użytkownika na podstawie interakcji z systemem opartym o technologie WWW*, praca magisterska, Uniwersytet im. Adama Mickiewicza w Poznaniu.

George, C., Scerri, J., 2007, *Web 2.0 and User-Generated Content: Legal Challenges in the New Frontier*, Journal of Information, Law and Technology, vol. 2.

Govani, T., Pashley, H., 2007, *Student Awareness of the Privacy Implications when Using Facebook*, Unpublished manuscript retrieved, September.

Google. Raport przejrzystości, 2012, http://www.google.com/governmentrequests/overview.html [dostęp 17.04.2012].

Gray, M.J., 2009, *Applying Nuisance Law to Internet Obscenity*, Journal of Law and Policy for the Information Society, vol. 6, no. 2, s. 317.

Guarino, N., 1998, *Formal Ontology and Information Systems*, IOS Press, s. 3–15.

Henze, N., Nejdl, W., 2003, *Logically Characterizing Adaptive Educational Hypermedia Systems*, w: *In International Workshop on Adaptive Hypermedia and Adaptive Web-based Systems* (*AH 2003*), 20–24 May 2003, Budapest, Hungary.

Hypponen, M., 2007, *Online Crime and Crime Online*, http://video.google.com/videoplay?docid=-4310105531136655049, http://conference.hitb.org/hitbsecconf2007kl/materials/, prezentacja (keynote 2) wygłoszona na konferencji HITBSecConf2007: Deep Knowledge Security Conference w Malezji, 5 września 2007.

Kaczmarek, T., 2006, *Integracja danych z głębokiego Internetu dla potrzeb analizy otoczenia przedsiębiorstwa*, rozprawa doktorska, Akademia Ekonomiczna w Poznaniu.

Koch, N., 2000, *Software Engineering for Adaptive Hypermedia Systems*, rozprawa doktorska, Ludwig-Maximilians-Universität München.

Królikowski, P., 2009, *Dowody elektroniczne – problematyka zabezpieczania i dobre praktyki*, w: Jemioła T., Kisielnicki J., Rajchel K. (red.), *Cyberterroryzm – nowe wyzwania XXI wieku*, Wyższa Szkoła Informatyki, Zarządzania i Administracji, Warszawa, s. 371–395.

Liu, H., Azadegan, S., Yu, W., Acharya, S., Sistani, A., 2012, *Are We Relying Too Much on Forensics Tools?*, w: *Software Engineering Research, Management and Applications 2011*, vol. 377, Springer, Berlin-Heidelberg, http://link.springer.com/book/10.1007/978-3-642-23202-2/page/1, s. 145–156.

Manning, C.D., Raghavan, P., Schütze, H., 2008, *Introduction to Information Retrieval*, Cambridge University Press.

McKemmish, R., 2008, *When Is Digital Evidence Forensically Sound*, w: Ray, I., Shenoi, S. (eds.), *Advances in Digital Forensics IV, Fourth Annual IFIP WG 11.9 Conference on Digital Forensics, Kyoto University, Kyoto, Japan, January 28–30, 2008*, Springer, s. 3–15.

Miller, C., 2010, *Investigating the Social Web: New Training Helps Investigators Find Criminals and Evidence on the Internet*, Law Enforcement Technology, vol. 37, no. 1, s. 6.

Mukasey, M.B., Sedgwick, J.L., Hagy, D.W., 2008, *Electronic Crime Scene Investigation — A Guide for First Responders, Second Edition*, Tech. Rep., U.S. Department of Justice, National Institute of Justice.

Pareja, S., 2008, *Sales Gone Wild: Will the FTC's Business Opportunity Rule Put an End to Pyramid Marketing Schemes*, McGeorge L. Rev., vol. 39, s. 83.

Pollitt, M., 2008, *Applying Traditional Forensic Taxonomy to Digital Forensics*, w: Ray, I., Shenoi, S. (eds.), *Advances in Digital Forensics IV, Fourth Annual IFIP WG 11.9 Conference on Digital Forensics, Kyoto University, Kyoto, Japan, January 28–30, 2008*, Springer, s. 17–26.

Rajagopalan, S., 2000, *A Study of Security Problems Associated with the Telephone Network*, Tech. Rep., Oregon State University, Department of Electrical and Computer Engineering.

Rush, H., Smith, C., Kraemer-Mbula, E., Tang, P., 2008, *Crime Online: Cybercrime and Illegal Innovation*, Tech. Rep.

Solomon, M., Broom, N., Barrett, D., 2004, *Computer Forensics JumpStart*, SYBEX Inc., Alameda, CA, USA.

Sowa, J.F., 2006, *Semantic Networks*, http://www.jfsowa.com/pubs/semnet.htm [dostęp: 18.04.2012].

US-CERT, 2008, *Computer Forensics*, Tech. Rep., US-CERT.

Wahlster, W., Kobsa, A., 1989, *User Models in Dialog Systems*, Springer.

Wall, D.S., 2005, *The Internet as a Conduit for Criminals*, Cage, Thousand Oaks, CA, s. 77–98.

Whitty, M., Young, G., Goodings, L., 2011, *What I Won't Do in Pixels: Examining the Limits of Taboo Violation in MMORPGs*, Computers in Human Behavior, vol. 27, no. 1, s. 268–275.

PERSONAL INFORMATION RETRIEVAL FOR PURPOSES OF COMPUTER FORENSICS

Abstract: Committing a crime with the use of the Internet infrastructure is easier and involves less risk than committing a crime in the traditional way. Fighting crime in the cyberspace is possible due to footprints that each user leaves. This article discusses the problem of crime in the cyberspace, and issues related to combating such crime. A method is presented to search for the users' personal information by extending available information about them. Additionally, empirical evaluation of the effectiveness of this method is performed.