Zapewnienie bezpieczeństwa przez semantyczne monitorowanie cyberprzestrzeni

Witold Abramowicz, *Elżbieta Bukowska*, *Agata Filipowska*

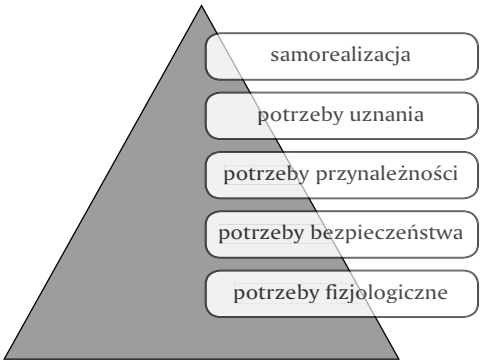
Bezpieczeństwo jest jedną z podstawowych potrzeb człowieka. Wraz ze zwiększaniem się aktywności społeczeństwa w cyberprzestrzeni pojawia się również silniejsza potrzeba zapewnienia bezpieczeństwa w tym obszarze. Zagrożenia manifestujące się w internecie nie ograniczają się jedynie do wirtualnej przestrzeni – mogą oddziaływać na fizyczny świat. Przykładami takich zagrożeń są ogłoszenia sprzedaży nielegalnych towarów lub próby wyłudzeń. Szybkie tempo rozwoju technologicznego powoduje, że również tradycyjne metody pozyskiwania informacji stają się niewystarczające do automatycznego monitorowania sieci w celu wykrycia tego typu potencjalnych zagrożeń.

Artykuł przedstawia problem monitorowania źródeł w cyberprzestrzeni oraz wskazuje na istniejącą lukę pomiędzy wolumenem dostępnych informacji a zdolnością do podejmowania decyzji przez jednostki odpowiedzialne za ocenę zagrożenia w internecie (np. policję, administratorów i moderatorów stron). Prezentuje też jedną z inicjatyw z zakresu zapewnienia bezpieczeństwa w cyberprzestrzeni, jaką jest projekt „Semantyczny Monitoring Cyberprzestrzeni".

Piramida potrzeb Maslowa[1] ukazuje, że ludzie chcą zaspokajać potrzeby związane z bezpieczeństwem zaraz po potrzebach fizjologicznych (por. rysunek 1). Według antropologii społecznej człowiek nauczył się zaspokajać swoje potrzeby bezpieczeństwa we współpracy z innymi.

Współcześnie potrzeby te zaspokajane są w różnorodny sposób: człowiek dba o swoje bezpieczeństwo samodzielnie, ale także we współdziałaniu z rodziną, grupami formalnymi i nieformalnymi. Szereg zadań związanych z zapewnieniem bezpieczeństwa przyjęły na siebie instytucje państwa i samorządu terytorialnego. Bezpieczeństwo dotyczy nie tylko poszczególnych osób, ale także grup społecznych, narodów i społeczności międzynarodowej[2]. Odnosi się do najbliższego otoczenia każdego człowieka – określonych terenów, ale także do całych państw oraz ich grup. Jest wyrazem polityki regionalnej, polityki państw oraz sensem istnienia sojuszy międzynarodowych. Poziom bezpieczeństwa

Rysunek 1. Piramida potrzeb według Abrahama Maslowa



Źródło: A.H. Maslow, Motivation and Personality, Pearson 1997, tłumaczenie własne

można mierzyć w odniesieniu do dowolnego miejsca w przestrzeni fizycznej. Bywają dzielnice miast, które są bezpieczniejsze, i takie, które są mniej bezpieczne. Może on też być zmienny w czasie. Dzielnice bywają bezpieczne w dzień i niebezpieczne w nocy. Liban, nazywany kiedyś „Szwajcarią Bliskiego Wschodu", teraz jest krajem podwyższonego ryzyka. Państwa zbudowały organizacje wyspecjalizowane w zapewnieniu bezpieczeństwa, np. policję, wojsko, straż pożarną.

Powszechne zastosowanie systemów informacyjnych stworzyło nowy obszar, w którym chcemy czuć się bezpiecznie – cyberprzestrzeń. Pojęcie „cyberprzestrzeń" zawdzięczamy Stanisławowi Lemowi, który użył go po raz pierwszy w 1964 roku[3]. Współczesne jego znaczenie związane jest z upowszechnieniem internetu, stwarzającym wspólną przestrzeń udostępniania danych i informacji. Cyberprzestrzeń jest często zupełnie „odmiejscowiona", czasami ma jednak jakiś związek z przestrzenią fizyczną. Zapewnienie bezpieczeństwa w cyberprzestrzeni znacznie wykracza poza doświadczenia związane z bezpieczeństwem w świecie rzeczywistym. Uczenie się, jak zapewnić sobie bezpieczeństwo w przestrzeni fizycznej zabrało ludziom wieki. Cyberprzestrzeń powstała w ciągu

[1] A.H. Maslow, *Motivation and Personality*, Pearson, 1997.

[2] S. Koziej, *Wstęp do teorii i historii bezpieczeństwa*, Warszawa 2010, http://www.koziej.pl/index.php?pid=5, [13.06.2013].

[3] S. Lem, *Summa technologiae*, Wydawnictwo Literackie, Kraków 1964. Wielu autorów piszących o cyberprzestrzeni pierwsze użycie tego pojęcia przypisuje jednak Williamowi Gibsonowi, który opublikował w 1984 roku *Burning Chrome*.

zaledwie kilku ostatnich dziesięcioleci. Tak szybko przystosowywać się do zmian społecznych nie potrafią ani ludzie, ani instytucje państwowe. Także ze względu na pozycję potrzeb bezpieczeństwa w piramidzie Maslowa zapewnienie bezpieczeństwa w cyberprzestrzeni jest ogromnym wyzwaniem.

Źródła danych i informacji w cyberprzestrzeni

Powszechnie uważa się, że największe zasoby informacji w cyberprzestrzeni dostępne są za pośrednictwem wyszukiwarek internetowych. Nic bardziej mylnego. Tę część internetu nazywamy internetem widocznym (*visible web, indexable web lub surface web*) dla odróżnienia od internetu głębokiego (*deep web, hidden web, invisible web*), którego zasoby nie są osiągalne w opisany sposób[4]. Przyczyny niemożności dotarcia do nich mogą być różne:

1. Dostęp do danych możliwy jest przez znalezienie odpowiedzi w bazie danych po sformułowaniu zapytania.
2. Brak powiązań źródła głębokiego internetu z innymi źródłami uniemożliwia nawigację do takiego źródła, w konsekwencji roboty wyszukiwarek internetowych nie mogą do niego dotrzeć.
3. Wiele zasobów głębokiego internetu dostępnych jest po spełnieniu dodatkowych warunków wykorzystania zasobów, np. po uprzednim zarejestrowaniu się lub podaniu hasła albo adresu IP.
4. Forma danych może przyczynić się do ich niedostępności w widocznym internecie – dotyczy to np. danych opisanych w innym formacie niż HTML lub aktywowanych przez skrypty, m.in. JavaScript lub Flash.

W literaturze można znaleźć różne poglądy na temat tego, jaką część sieci stanowi internet widoczny. Najbardziej optymistyczne dla wyszukiwarek internetowych dane wskazują, że jest to kilka procent zasobów internetu.

Źródła danych w internecie mogą być traktowane – w uproszczeniu – jako rozrastające się w czasie ogromne kolekcje wzajemnie powiązanych dokumentów. W takich zasobach można wyszukiwać i nawigować po nich lub je wertować[5]. Od kolekcji dokumentów odróżniamy strumienie danych. Są to np. dokumenty lub – ogólniej – dane, które mogą być przechwycone po ich udostępnieniu. Jednak gromadzenie i publikowanie danych w celu późniejszego ich wykorzystania nie jest intencją twórców (lub nie jest to ich najważniejsza intencja). Przykładem strumieni danych są pozycje bieżące samolotów dostępne na portalach internetowych[6].

Interesującym w kontekście bezpieczeństwa, choć leżącym poza przedmiotem tej publikacji zagadnieniem, jest monitorowanie informacji pochodzących z sensorów mierzących takie parametry, jak temperatura, wilgotność, warunki oświetlenia, w tym nasłonecznienia, poziom hałasu, właściwości obiektów materialnych. Monitorowanie może dotyczyć zjawisk o znaczeniu wojskowym (jednostek wojska, sprzętu i wyposażenia), zjawisk w środowisku naturalnym, (np. symptomów wskazujących, że może wystąpić trzęsienie ziemi lub powódź), przestrzeni powietrznej i kosmicznej (np. przewidywanie kolizji z obiektami kosmicznymi). Zarówno kolekcje, jak i strumienie mogą być dostępne w całości lub części w internecie głębokim lub widocznym.

Ważną częścią internetu są również treści tworzone przez społeczności. Ich przykłady można znaleźć w powszechnie wykorzystywanych portalach społecznościowych, takich jak Facebook czy Twitter, a także na forach społecznościowych. Portale społecznościowe mogą być częścią internetu widocznego lub głębokiego, mogą tworzyć kolekcje lub strumienie.

Zagrożenia bezpieczeństwa w cyberprzestrzeni

Korzystanie z każdego z przedstawionych rodzajów źródeł internetowych związane jest z określonym zagrożeniem[7]. Skala tych zagrożeń wykazuje tendencję rosnącą[8]. Przedmiotem niniejszego artykułu nie jest jednak ich analiza.

W opracowaniu autorzy skupiają się na analizie źródeł internetowych w celu pozyskania z cyberprzestrzeni informacji o zagrożeniach bezpieczeństwa bez względu na to, czy dotyczą one bezpieczeństwa przestrzeni fizycznej, czy też cyberprzestrzeni. Każde z opisanych źródeł może zawierać informacje o zagrożeniach bezpieczeństwa. Mogą się one odnosić do dowolnej sfery życia.

Zagrożenie bezpieczeństwa w tradycyjny sposób kojarzone jest z zagrożeniem militarnym. Ostatnie groźby Korei Północnej są ilustracją zagrożenia, które może być postrzegane jako zagrożenie szantażem militarnym, prowokacja, incydent graniczny lub użycie środków przemocy zbrojnej. Przykład ten pokazuje bogactwo problemów do rozwiązania: należy znaleźć odpowiednie źródła, ze źródeł tych pozyskać informacje o potencjalnych zagrożeniach, dokonać oceny jakości zdobytych informacji (w zakresie takich miar jak aktualność, wiarygodność, retrospektywność, predykcja, użyteczność, zupełność, przyswajalność, a także relewancja kognitywna, czasowa, osobowa, lokalizacyjna i ekonomiczna[9]).

[4] W. Abramowicz, *Filtrowanie informacji*, Wydawnictwo Akademii Ekonomicznej w Poznaniu, 2008.
[5] Tamże.
[6] Flightradar24, http://www.flightradar24.com/, [13.06.2013].
[7] *Internet Security Threat Report 2013*, t. 18, kwiecień 2013, Symantec Corporation.
[8] *Internet Security Threat Report, Appendix 2013*, t. 18, kwiecień 2013, Symantec Corporation.
[9] W. Abramowicz, dz.cyt.

Cyberprzestrzeń dostarcza bogatych informacji o zagrożeniach ekonomicznych. Mogą być one rozważane w kategoriach makroekonomicznych lub mikroekonomicznych. W tym pierwszym przypadku analiza informacji w cyberprzestrzeni pozwala identyfikować i oceniać zagrożenia dotyczące np. realizacji polityki pieniężnej czy fiskalnej. Ilustracją tej ostatniej kategorii zagrożeń jest badanie polskiego rynku handlu prętami stalowymi, które ułatwia identyfikowanie skali wyłudzeń VAT, ale pozwala także na rozpoznanie konkretnych działań przestępczych. W zakresie identyfikowania zagrożeń mikroekonomicznych dzięki analizie informacji w cyberprzestrzeni możliwe jest np. szacowanie zagrożeń wynikających z działań konkurencji, dostawców lub klientów. Analiza portali społecznościowych pozwala m.in. na wykrywanie oszustw w zakresie nieuprawnionej likwidacji szkód ubezpieczeniowych. Dotyczy to zarówno poznawania nowych sposobów oszukiwania, jak i identyfikacji konkretnych nieuczciwych klientów firm ubezpieczeniowych.

Zagrożenia społeczne mogą także być rozpoznawane poprzez analizę cyberprzestrzeni. Dzięki niej możliwe jest prowadzenie odpowiedniej prewencyjnej polityki społecznej, np. w zakresie przeciwdziałania uprzedzeniom kulturowym lub ksenofobii. Możliwe jest także identyfikowanie konkretnych deliktów związanych z określonymi patologiami społecznymi, np. aktów pedofilii.

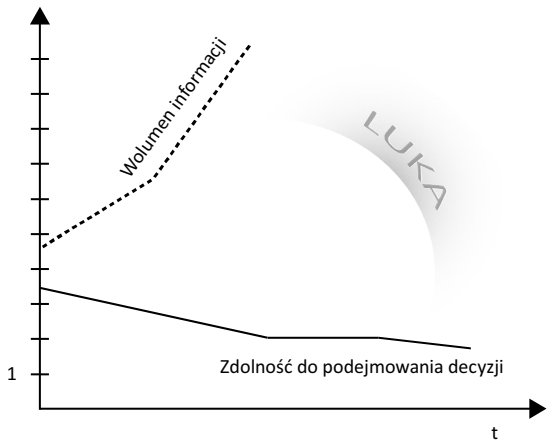
Wszystkie analizowane kategorie zagrożeń mogą dotyczyć zarówno bezpieczeństwa wewnętrznego, jak i zewnętrznego. W przypadku wszystkich można też mieć do czynienia z wynikami teoriomnogościowymi. Rozpoznawane są poszczególne zbiory zagrożeń – ocenia się ich moc oraz relacje pomiędzy nimi[10]. Można także dokonywać identyfikacji konkretnych zagrożeń – realizacja tego celu jest przedmiotem dalszych rozważań.

Monitorowanie cyberprzestrzeni identyfikujące zagrożenia

Identyfikacja konkretnych zagrożeń związana jest z monitorowaniem różnych źródeł internetu widocznego i głębokiego. Przez monitorowanie w opisywanym znaczeniu należy rozumieć proces pozyskiwania informacji o zagrożeniach możliwie szybko po ich opublikowaniu w źródle. Bardzo trudne – ale też z punktu widzenia osiągania celu nieuzasadnione – jest monitorowanie całego internetu. Dla każdej opisanej uprzednio kategorii zagrożeń i dla konkretnych zagrożeń z określonej kategorii należy zidentyfikować źródła, które powinny być monitorowane, aby pozyskać możliwie dużo informacji. Określenie takich źródeł jest działaniem eksperckim. Ekspert powinien ocenić nie tylko, z jakim prawdopodobieństwem źródło może informować o zagrożeniach, ale także jaka jest jakość źródła. W tym zakresie stosuje się podobne miary jak w przypadku jakości informacji[11]. Zakłada się, że źródła wysokiej jakości dają także informacje wyższej jakości.

Drugim argumentem przemawiającym za pozyskiwaniem informacji z dedykowanych źródeł jest przeciwdziałanie „luce decyzyjnej" (rysunek 2). Po przekroczeniu pewnego wolumenu informacji o zagrożeniach obniża się zdolność ich wykorzystania. Dlatego dla podwyższenia skuteczności monitorowania każdemu zagrożeniu można przypisać ryzyko z nim związane. Ryzyko pozwala szeregować informacje o zagrożeniach. Kryteria szeregowania mogą być jednak różnorodne. Pierwsza grupa wynika z przyporządkowania zagrożeniu miar informacji o nim. Na przykład ważniejsze może być zagrożenie opisane informacją aktualniejszą i bardziej wiarygodną niż mniej aktualną i mniej wiarygodną. Te miary szeregowania zagrożeń nie powinny występować jednak samodzielnie. Druga grupa miar szeregowania odnosi się bezpośrednio do oceny zagrożenia, a nie informacji o nim. Taką miarą jest potencjalna szkodliwość zagrożenia, np. informacja o bombie podłożonej przy mało uczęszczanej drodze powinna być w rankingu informacji o zagrożeniach umieszczona niżej niż informacja o takiej samej bombie umieszczonej przy drodze, na której zorganizowany jest masowy maraton. Ten sam przykład ilustruje również inną miarę zagrożeń, jaką jest oddziaływanie społeczne. Bardzo trudna do oszacowania jest miara prawdopodobieństwa zagrożenia, np. do początku tego wieku prawdopodobieństwo wystąpienia aktu terroryzmu związanego ze świadomym uderzeniem samolotu pasażerskiego w wieżowiec oceniane było jako znikome.

Rysunek 2. Luka pomiędzy wolumenem informacji a zdolnością do podejmowania decyzji[12]



Źródło: W. Abramowicz, P.J. Kalczyński, K. Węcel, Filtering the Web to Feed Data Warehouses, Springer Verlag, Londyn 2002

[10] Tamże.
[11] J. Oleński, *Ekonomika informacji. Metody*, PWE, Warszawa 2003.
[12] W. Abramowicz, dz.cyt., [za:] K. Węcel, *Tworzenie profili hurtowni danych do filtrowania informacji ekonomicznej*, Wydział Ekonomii, Akademia Ekonomiczna w Poznaniu (praca doktorska).

Pierwsza grupa miar wynika z analizy infologicznej zagrożeń. Dla oceny tych miar zastosowano kryteria znane z systemów wyszukiwawczych, takie jak pełność, precyzja i odpad. Druga grupa związana jest przeprowadzeniem analizy semantycznej informacji o zagrożeniach. Ocena skuteczności analizy semantycznej jest bardziej wiarygodna, jeżeli uwzględnia przedmiotowość zagrożeń, np. monitorowanie środowiska naturalnego.

Przeciwstawną miarą do ryzyka wystąpienia zagrożenia jest ryzyko wynikające z pominięcia informacji o zagrożeniu.

Monitorowanie cyberprzestrzeni w celu identyfikacji zagrożeń może być związane z dwoma paradygmatami pozyskiwania informacji: wyszukiwaniem i filtrowaniem[13]. W przypadku wyszukiwania mamy do czynienia z jednym zapytaniem i wieloma dokumentami. Filtrując, analizujemy każdorazowo jeden dokument i staramy się zaspokoić potrzeby informacyjne wszystkich użytkowników filtru. W trakcie wyszukiwania analizujemy więc zbiory dokumentów, w trakcie filtrowania przetwarzamy strumienie dokumentów.

Filtrowanie informacji daje aktualniejsze wiadomości o zagrożeniach, ponieważ przeprowadzane jest na bieżąco na strumieniu informacji. Trudniej w tym przypadku zastosować takie miary, jak wspomniane pełność, precyzja i odpad.

Wyszukiwanie informacji wymaga określenia periodyzacji odwiedzania źródeł internetowych. Utrudnia to pozyskanie aktualnych informacji.

Semantyczne monitorowanie cyberprzestrzeni

Ze względu na wolumen danych pojawiających się w internecie manualne przeprowadzanie procesu wyszukiwania lub filtrowania w celu identyfikacji potencjalnych zagrożeń jest nieefektywne. Zastosowanie w tym zakresie znajdują metody automatycznego monitorowania sieci. Najprostszym rozwiązaniem jest automatyczne wyszukiwanie na podstawie słownika słów kluczowych. Taki słownik powinien być dostosowany do domeny oraz na bieżąco aktualizowany. Wyszukiwanie poprzez słowa kluczowe zwraca jednak wiele niedopasowanych wyników.

Zyskujące coraz większą popularność technologie semantyczne uzupełniają wyszukiwanie słownikowe o zastosowanie specjalnych reguł wnioskowania, przez co umożliwiają bardziej efektywne, automatyczne monitorowanie źródeł i wyszukiwanie informacji odpowiadających zadanym kryteriom.

Wyzwanie stanowi zdefiniowanie reguł automatycznego przetwarzania, które umożliwią uzyskanie odpowiednich miar precyzji i pełności zwracanych wyników, tym samym zmniejszając lukę decyzyjną w procesie oceny zagrożenia. W przypadku zagrożeń w cyberprzestrzeni największym problemem jest „zrozumienie" przez system ich treści, w tym uwzględnienie stosowanych kolokwializmów, skrótów, błędów pisowni. Ponadto zarówno źródła, jak i treści ogłoszeń charakteryzują się dużą dynamiką zmian i są często przedmiotem moderacji.

Wśród wielu inicjatyw zajmujących się kwestią eliminowania zagrożeń pojawiających się w cyberprzestrzeni wymienić można projekt *Semantyczny Monitoring Cyberprzestrzeni* (SMC)[14]. Głównym celem projektu było opracowanie prototypu narzędzia, które pozwoli na ciągłe monitorowanie wybranych przez eksperta źródeł internetowych dla identyfikacji pojawiających się w nich określonych aktywności mogących wskazywać na działania przestępcze. Ograniczenie czynności do z góry zdefiniowanej grupy źródeł pozwala monitorować je z większą częstotliwością, a przez to zwiększyć szanse na pozyskanie aktualnych danych. Ponadto, z uwagi na zróżnicowanie źródeł pod względem treści i technologii, wybranie ich podzbioru umożliwia stworzenie precyzyjnych reguł ekstrakcji, które pozwalają na pozyskanie danych spoza właściwej treści dokumentu (np. ze znaczników HTML, nazw klas, niewyświetlanych pól). Ogłoszenia dotyczące nielegalnej działalności pojawiają się głównie na słabo moderowanych stronach, gdzie możliwe jest anonimowe dodawanie treści. Zatem ograniczenie zakresu wyszukiwania do tych stron przyczynia się do zwiększenia efektywności narzędzia.

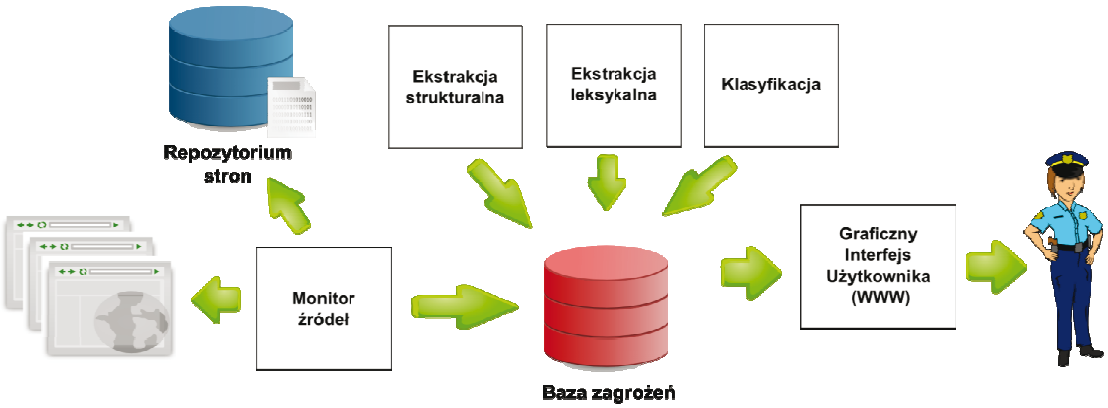
Projekt SMC

Z uwagi na duże zróżnicowanie zagrożeń w cyberprzestrzeni, które zostało omówione wcześniej w artykule, prototyp systemu SMC został opracowany dla jednego ich typu, z zachowaniem uniwersalności rozwiązania, tak aby możliwe było przeniesienie go na inne domeny. Od początku realizacji projektu planowane było w celach testowych wdrożenie prototypu SMC w jednostce Komendy Głównej Policji. Po konsultacjach z odbiorcą jako przypadek użycia wybrano nielegalny handel lekami i produktami medycznymi. Zgodnie z polskim prawem niedopuszczalny jest handel lekami dostępnymi na receptę lub wycofanymi z obrotu. Leki dostępne bez recepty mogą natomiast być sprzedawane za pośrednictwem internetu wyłącznie przez uprawnione do tego apteki, poprzez ich własne strony internetowe spełniające określone ustawą warunki. Niekontrolowany handel lekami (tzn. przeprowadzony poza obrotem aptecznym) może powodować zagrożenie życia i zdrowia osoby, która przyjmie dany lek. Medykamenty pochodzące z nielegalnego obrotu mogą być podrabiane, przechowywane w nieodpowiednich warunkach, źle zabezpieczone w czasie transportu lub sprzedawane po upływie

[13] W. Abramowicz, dz.cyt.

[14] Projekt zrealizowany przez Katedrę Informatyki Ekonomicznej, Uniwersytet Ekonomiczny w Poznaniu oraz Sygnity SA, finansowany przez Narodowe Centrum Badań i Rozwoju (nr kontraktu 0079/R/T00/2010/11). Strona projektu: http://smc.kie.ue.poznan.pl.

Rysunek 3. Komponenty systemu SMC



Źródło: T. Kaczmarek, J. Dzikowski, Architektura systemu wykrywania zagrożeń w cyberprzestrzeni, konferencja „Technologie informatyczne w administracji publicznej i służbie zdrowia", 08.12.2011 r., Warszawa

terminu ich ważności. Osobnym problemem, leżącym poza obszarem badań projektu SMC, jest motywacja osób, które bez wcześniejszej konsultacji z lekarzem decydują się na przyjmowanie leków dostępnych tylko na receptę lub wycofanych z obrotu.

System SMC zbudowany jest z modułów (rysunek 3)[15] i operuje na tzw. profilach zagrożeń. Profil zagrożenia gromadzi informacje o jednej ofercie kupna, sprzedaży lub wymiany opublikowanej w sieci, w tym informacje o przedmiocie wymiany (lub przedmiotach), autorze ogłoszenia oraz dane związane z czasem i miejscem publikacji. Każde ogłoszenie, jakie pojawi się w wybranych źródłach, posiada w systemie SMC swoją reprezentację w postaci profilu zagrożenia.

Ponieważ sprzedaż leków przez internet (poza uprawnionymi do tego aptekami) jest prawnie zabroniona, większość źródeł na bieżąco usuwa tego typu ogłoszenia. Z drugiej strony, z uwagi na liczbę ogłoszeń i brak narzędzi do automatycznej analizy ich treści, proces moderacji trwa zwykle od kilku godzin do kilku dni. Jeżeli ogłoszenie zawierające ofertę nielegalnej sprzedaży zostanie znalezione przez system, konieczne jest wykonanie jego kopii w celu umożliwienia ewentualnego śledztwa. Problem pozyskiwania kopii ogłoszeń został rozwiązany poprzez opracowanie modułu *Monitor źródeł*. Robot internetowy odwiedza systematycznie zdefiniowane źródła i pobiera do bazy danych kopie wszystkich podstron, które się w nich znajdują. W czasie ponownej wizyty w źródle pobiera już tylko nowe i zmienione podstrony.

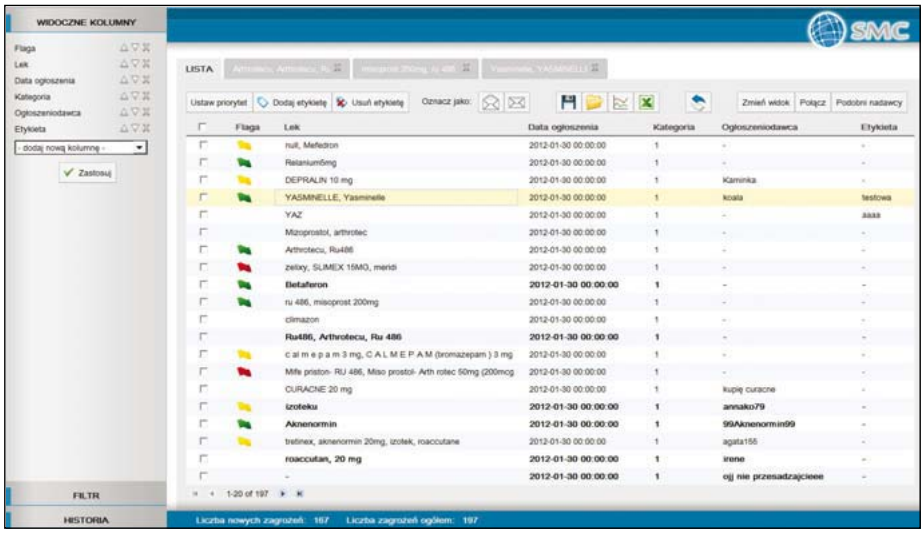
Przetwarzanie danych odbywa się w dwóch etapach: najpierw przeprowadzana jest analiza struktury źródła (moduł ekstrakcji strukturalnej), a następnie semantyczna analiza treści samego ogłoszenia (moduł ekstrakcji leksykalnej). Źródła internetowe są bardzo zróżnicowane pod względem technicznym: posty na forach mają temat i nazwę użytkownika zapisane w osobnych polach, niektóre portale ogłoszeń stosują specjalne pole dla oznaczenia lokalizacji sprzedającego, pola zawierające te same informacje mogą mieć różne identyfikatory („użytkownik", „nazwa użytkownika", „imię"). Z tego względu konieczne było zdefiniowanie zestawu reguł ekstrakcji strukturalnej odpowiadających strukturze każdego ze zdefiniowanych źródeł. W wyniku ekstrakcji stworzony zostaje zbiór dokumentów logicznych. Jedna strona pobrana przez robota internetowego może zostać podzielona na kilka dokumentów logicznych, z których każdy odpowiada jednemu ogłoszeniu, postowi na forum lub komentarzowi. Efektem działania ekstrakcji strukturalnej jest częściowe uzupełnienie profilu zagrożenia oraz przygotowanie dokumentów logicznych.

Za analizę semantyczną treści ogłoszeń odpowiada moduł ekstrakcji leksykalnej. Wynikiem jego działania są adnotacje wskazujące przedmiot obrotu (nazwy leków wraz z oferowaną ilością, dawką, typem opakowania, datą ważności lub informacją o stanie produktu) oraz o podmiocie oferującym sprzedaż lub zakup. Moduł korzysta z zestawu leksykonów odpowiadających domenie oraz zasad adnotacji poszczególnych części ogłoszeń zapisanych jako formalne reguły, które można automatycznie przetwarzać. Adnotacje stworzone przez moduł ekstrakcji leksykalnej uzupełniają profile zagrożeń przechowywane w bazie danych. Adnotacja nazw leków uwzględnia synonimy lub często pojawiające się błędy w pisowni. Adnotacja typu akcji (kupno, sprzedaż) wykorzystuje reguły uwzględniające specyfikę języka polskiego i wieloznaczność niektórych określeń.

Kolejnym elementem jest automatyczna klasyfikacja profili umożliwiająca identyfikację ogłoszeń zawierających zagrożenia. Wiele dokumentów logicznych zebranych przez monitor źródeł zawiera oferty sprzedaży innych produktów niż leki (np. tabletki do zmywarek, kosmetyki), informacje o sprzedawcach (np. ostrzeżenia przed oszustami albo rekomendacje) lub dodatkowe informacje na temat leku (np. jakie wystąpiły skutki uboczne lub jak należy go

[15] T. Kaczmarek, J. Dzikowski, *Architektura systemu wykrywania zagrożeń w cyberprzestrzeni*, konferencja *Technologie informatyczne w administracji publicznej i służbie zdrowia*, 08.12.2011 r., Warszawa.

Rysunek 4. *Semantyczny Monitoring Cyberprzestrzeni* – interfejs prototypu

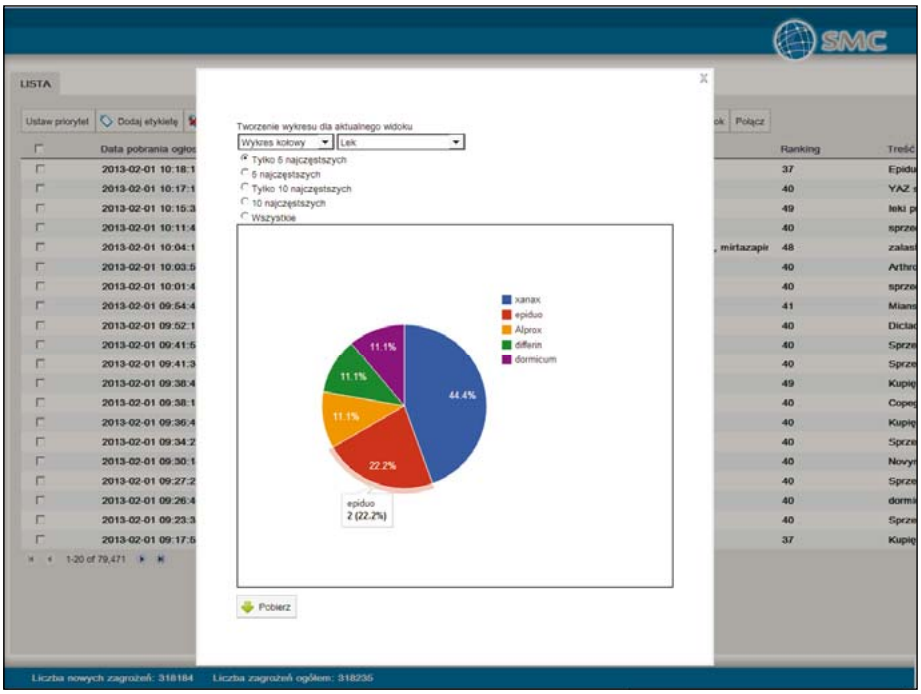


Źródło: opracowanie własne

dawkować). Dokumenty te mogą służyć jako źródła dodatkowe, objaśniające zagrożenia, ale nie powinny być klasyfikowane jako zagrożenie. Moduł klasyfikacji ocenia prawdopodobieństwo, że dany profil rzeczywiście reprezentuje zagrożenie nielegalnym handlem lekami. Podczas klasyfikacji wykorzystywane są miary otrzymane w wyniku analizy infologicznej oraz analizy semantycznej. Ponadto system przyznaje wyższą ocenę ogłoszeniom, które dotyczą większej liczby leków lub większej ilości leku. Ocena zależy również m.in. od liczby ogłoszeń i częstotliwości ich publikowania przez tego samego sprzedawcę.

System wymaga opracowania interfejsu, który pozwoli użytkownikowi końcowemu efektywnie wyszukiwać i analizować zebrane profile zagrożeń. Moduł GUI (*Graphical User Interface*) został zaimplementowany jako aplikacja sieciowa, dostępna przez przeglądarkę internetową (rysunek 4). Użytkownik ma możliwość szeregowania profili zagrożeń według daty publikacji lub nadanej przez system oceny. Możliwe jest również filtrowanie według dowolnych elementów profilu, zapisywanie parametrów filtra oraz generowanie statystyk (rysunek 5). Dla każdego profilu zagrożenia dostępny jest podgląd szczegółowy, który zawiera odnośnik do dokumentu logicznego przechowywanego w repozytorium oraz link do oryginalnego ogłoszenia. Jako uzupełnienie GUI stworzono również aplikację mobilną, umożliwiają-

Rysunek 5. *Semantyczny Monitoring Cyberprzestrzeni* – funkcja generowania statystyk



Źródło: opracowanie własne

cą natychmiastowy dostęp do statystyk pobranych informacji o zagrożeniach.

Należy podkreślić, że system nie informuje o popełnieniu przestępstwa, a jedynie pomaga wyszukiwać ogłoszenia, które mogą stanowić zagrożenie dla ich odbiorcy (próby oszustwa lub nielegalnego handlu) i być przesłanką stwierdzenia możliwości popełnienia przestępstwa. Ostateczna ocena ogłoszenia może być dokonana wyłącznie przez eksperta (w tym przypadku policjanta), który decyduje o podjęciu dalszych czynności.

Podsumowanie

Kwestia bezpieczeństwa w cyberprzestrzeni jest stosunkowo nowym problemem. Gwałtowny rozwój internetu i upowszechnienie się jego wykorzystania przez społeczeństwo powoduje, że coraz większą aktywność w cyberprzestrzeni wykazują również grupy przestępcze. Zagrożenia manifestujące się w sieci mogą odnosić się do świata fizycznego. Takie zagrożenia mogą skutkować utratą życia lub zdrowia przez osobę, która padnie ich ofiarą (np. na skutek ataku terrorystycznego lub zażycia podrabianego leku). Metody powszechnie kojarzone z zapewnieniem bezpieczeństwa w sieci, np. programy antywirusowe, nie znajdują zastosowania w przypadku tej grupy zagrożeń. Co więcej, przestępcy w błyskawicznym tempie adaptują do swych potrzeb nowe technologie oraz wykorzystują istniejące luki prawne, przez co ich działalność jest trudna do wykrycia przy pomocy tradycyjnych metod pozyskiwania informacji.

Coraz częściej dostrzega się możliwość wykorzystania technologii semantycznych do prowadzenia automatycznego, ciągłego monitoringu cyberprzestrzeni w celu identyfikacji pojawiających się w niej zagrożeń. Projekt Semantyczny Monitoring Cyberprzestrzeni jest przykładem inicjatywy wykorzystującej technologie semantyczne jako uzupełnienie metod ekstrakcji informacji i przetwarzania tekstu dla zwiększenia bezpieczeństwa użytkowników internetu.

Przykład projektu SMC dowodzi, że istnieje potrzeba tworzenia automatycznych narzędzi wspomagających działania zmierzające do zwiększenia bezpieczeństwa w cyberprzestrzeni.

Witold Abramowicz jest profesorem ekonomii, kierownikiem Katedry Informatyki Ekonomicznej Uniwersytetu Ekonomicznego w Poznaniu. Kieruje również projektem SMC – Semantyczny Monitoring Cyberprzestrzeni, finansowanym przez Narodowe Centrum Badań i Rozwoju. Kierował licznymi projektami w ramach 6. i 7. Programu Ramowego UE oraz innych programów badawczych UE, a także w ramach projektów NCN i NCBiR. Jest autorem lub współautorem 36 książek, ponad 300 publikacji w czasopismach, rozdziałów w monografiach oraz publikacji konferencyjnych.

Agata Filipowska jest doktorem ekonomii i informatyki, kierownikiem Next Generation Internet Lab działającego w Katedrze Informatyki Ekonomicznej Uniwersytetu Ekonomicznego w Poznaniu oraz koordynatorem projektu SMC. Koordynowała wiele projektów 6. i 7. Programu Ramowego UE, innych programów badawczych UE oraz projektów NCBiR, kierowała zadaniami lub uczestniczyła w nich. Jest autorką i współautorką kilkudziesięciu publikacji w czasopismach, rozdziałów w monografiach oraz publikacji konferencyjnych.

Elżbieta Bukowska jest doktorantką w Katedrze Informatyki Ekonomicznej Uniwersytetu Ekonomicznego w Poznaniu. Jest uczestniczką projektu SMC, brała też udział w projektach NCN i NCBiR. Jest autorką lub współautorką kilkunastu publikacji w czasopismach, rozdziałów w monografiach oraz publikacji konferencyjnych.