Roczniki Kolegium Analiz Ekonomicznych
nr 29/2013

Krzysztof Węcel

Katedra Informatyki Ekonomicznej
Wydział Informatyki i Gospodarki Elektronicznej
Uniwersytet Ekonomiczny w Poznaniu

Jacek Piotrowski

Sygnity S.A.

Metoda oceny stopnia zagrożenia związanego z publikacją ogłoszeń internetowych

1. Wstęp

Jednym z istotnych osiągnięć w rozwoju Internetu jest jego demokratyzacja, a więc umożliwienie współtworzenia zawartości sieci przez osoby trzecie[1]. Wynika z tego wiele pozytywnych zjawisk (np. funkcjonowanie Wikipedii), ale nie można nie zauważyć również wielu negatywnych skutków. Poczucie anonimowości oraz zmniejszenie kontroli nad treścią powodują, że Internet stał się wygodną platformą zarówno dla tradycyjnych, jak i dla nowych rodzajów przestępstw: od publikacji obscenicznych treści poprzez nielegalny handel do kradzieży tożsamości.

Ochrona obywateli przed tego rodzaju przestępstwami jest szczególnym wyzwaniem dla państwa. Z jednej strony mamy do czynienia z treściami, które wyczerpują znamiona przestępstwa (np. pornografia dziecięca), a z drugiej – są to np. ogłoszenia, które wymagają dokładniejszego zbadania. Klasyfikacji przestępstw nie można dokonywać w sposób automatyczny. Wiele takich projektów w USA upadło ze względu na zarzuty o naruszenie prywatności oraz tworzenie zbyt wielu fałszywych podejrzeń[2].

---

[1] Nieposiadające własnych serwerów WWW, niebędące administratorami czy też nieznające HTML.

[2] T.E. Senator, *On the efficacy of data mining for security applications*, w: *Proceedings of the ACM SIGKDD Workshop on CyberSecurity and Intelligence Informatics – CSI-KDD '09*, ACM Press, New York 2009, s. 75–83, http://dl.acm.org/citation.cfm?id=1599272.1599286.

O tym, czy dany czyn jest przestępstwem, rozstrzyga sąd, co jest poprzedzone zebraniem odpowiednich dowodów. W każdym przypadku potrzebni są ludzie z odpowiednimi uprawnieniami oraz kompetencjami. Mogą jednak i powinni być wspierani przez odpowiednie narzędzia informatyczne pozwalające na monitoring cyberprzestrzeni oraz wychwytywanie relewantnych sygnałów. Każda sytuacja, która może prowadzić do przestępstwa, określana jest jako zagrożenie. Profil zagrożenia to informacje o zagrożeniu zebrane z danego źródła internetowego i ustrukturyzowane w postaci atrybutów oraz przypisanych im wartości, zgodnie z określoną klasą zagrożenia.

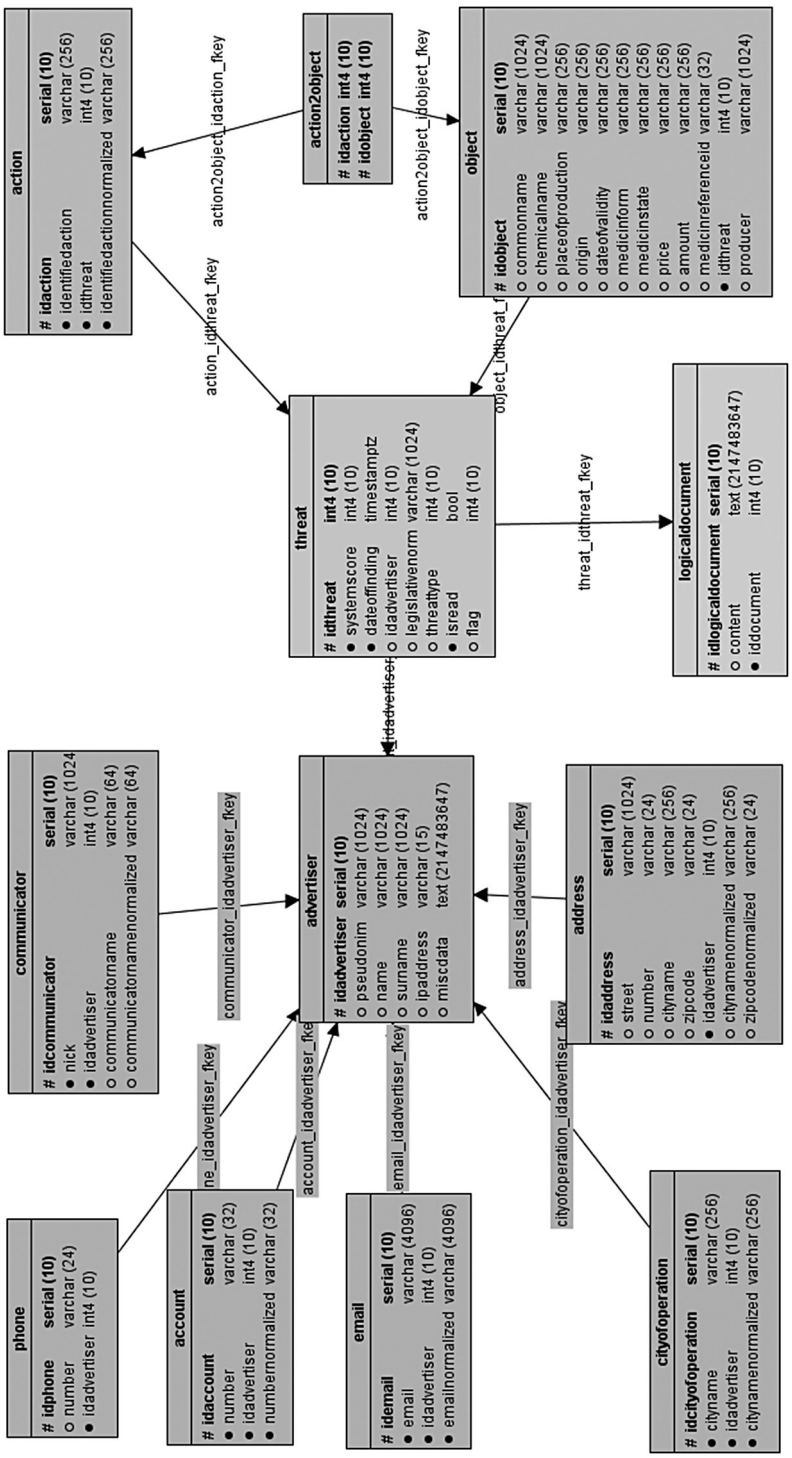
Celem niniejszego artykułu jest propozycja oceny stopnia zagrożenia pozwalająca na szeregowanie zagrożeń na potrzeby działań operacyjnych właściwych organów państwa.

2. Tło projektu oraz motywacja

Proponowana metoda znajduje zastosowanie w projekcie Semantyczny Monitoring Cyberprzestrzeni (SMC). Celem projektu jest opracowanie metody oraz prototypu narzędzia pozwalającego na przezwyciężenie problemu integracji danych i informacji pochodzących ze zróżnicowanych, ale uprzednio wskazanych źródeł na potrzeby ochrony cyberprzestrzeni poprzez wykrywanie zagrożeń, które manifestowane są w rozpatrywanych źródłach.

Na początku projektu została przeprowadzona szczegółowa analiza możliwych scenariuszy oraz potencjalnych zagrożeń. W jej wyniku określono, że najbardziej obiecującym obszarem zastosowań modelu SMC jest monitorowanie zagrożenia, jakim jest niezgodna z prawem sprzedaż leków inicjowana w Internecie. Zatem przedmiotem zainteresowania są pojedyncze ogłoszenia na forach internetowych, a proponowana metoda ma oceniać, na ile dane ogłoszenie stanowi zagrożenie dla porządku społecznego. Przypisana ocena pozwala na szeregowanie zagrożeń.

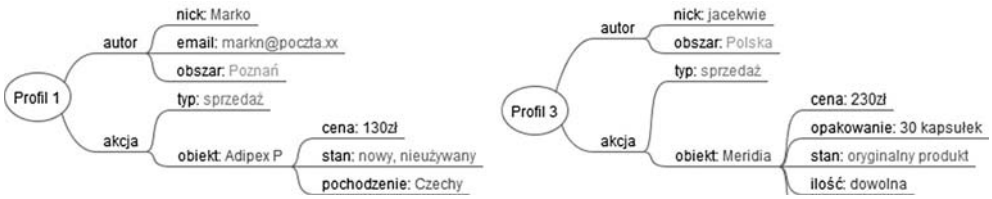
Dla zrozumienia rozważań związanych z metodą oceny zagrożeń konieczne jest przedstawienie samego profilu zagrożenia. Dla przyjętego scenariusza biznesowego został zdefiniowany schemat bazy danych, którego fragment dotyczący profilu zagrożenia został przedstawiony na rysunku 1.



Rysunek 1. Schemat profilu zagrożenia dla scenariusza nielegalnego obrotu lekami

Źródło: J. Małyszko, A. Filipowska, W. Abramowicz, T. Kaczmarek, E. Bukowska, B. Perkowski, P. Stolarski et al., *Architektura systemu wykrywania zagrożeń w cyberprzestrzeni*, „Roczniki" Kolegium Analiz Ekonomicznych SGH, z. 24, Oficyna Wydawnicza SGH, Warszawa 2012, s. 11–22.

Schemat przewiduje dość znaczną liczbę atrybutów. W rzeczywistych ogłoszeniach podane są wartości zaledwie kilku z nich. Rysunek 2 przedstawia przykładowe profile zagrożenia.



Rysunek 2. Przykładowe profile zagrożenia

Źródło: opracowanie własne.

3. Problem badawczy i prace powiązane

Zagadnienie klasyfikacji ogłoszenia jako zagrożenia jest szczególnym przypadkiem zagadnienia kategoryzacji tekstu. Istotnym wyzwaniem jest to, że nie ma gotowych rozwiązań dziedzinowych, a metody ogólne byłyby niewystarczające. Według naszej wiedzy nikt wcześniej nie podejmował prac w dziedzinie będącej przedmiotem zainteresowania projektu SMC.

Większość metod kategoryzacji tekstu opiera się na modelu wektorowym (*vector-space model*) znanym z wyszukiwania informacji[3]. Celem jest wychwycenie struktury takiej, jak kolejność i bliskość słów, fraz, ich lokalizacja w dokumencie.

We wszystkich podejściach można spotkać się z problemem redukcji przestrzeni. Na przykład Karras[4] proponuje metodykę ekstrakcji cech, która przede wszystkim wykorzystuje kategorie semantyczne zamiast surowych wystąpień słów. Ogura i współpracownicy[5] proponują nową miarę do selekcji cech, które przybliżają ważność danego termu, opierając się na tym, jak bardzo rozkład

[3] A. Markov, *Fast categorization of Web documents represented by graphs*, „Advances in Web Mining and Web Usage Analysis" 2007, vol. 4811, s. 56–71, http://www.springerlink.com/index/u4886005r4760437.pdf.

[4] D. Karras, *An improved text categorization methodology based on second and third order probabilistic feature extraction and neural network classifiers*, „Knowledge-Based Intelligent Information and Engineering System" 2006, vol. 4251, s. 9–20, http://www.springerlink.com/index/07m8415wj15v2677.pdf.

[5] H. Ogura, H. Amano, M. Kondo, *Feature selection with a measure of deviations from Poisson in text categorization*, „Expert Systems with Applications" 2009, vol. 36(3), s. 6826–6832, doi:10.1016/j.eswa.2008.08.006

prawdopodobieństwa każdego termu różni się od standardowego rozkładu Poissona. W dziedzinie wyszukiwania informacji odchylenie to jest wykorzystywane do nadawania wag słowom kluczowym. Innym pomysłem jest wykorzystanie ważenia TF-IDF, co również pozwala ograniczyć liczbę termów[6].

Przygotowana przestrzeń może być wykorzystana do uruchomienia właściwych algorytmów klasyfikacji. W literaturze można znaleźć próby wykorzystania Support Vector Machines (SVM)[7], Latent Semantic Indexing (LSI)[8] czy k-NN[9]. Przegląd technik automatycznej klasyfikacji tekstu znajduje się w pracy Sebastianiego[10], a jednym z ważniejszych zawartych w niej wniosków jest to, że naiwny klasyfikator Bayesa ze względu na swoją prostotę oraz przewagę w wydajności może być preferowany w stosunku do innych podejść. Jeśli chodzi o same metody klasyfikacji, to wydaje się, że ich efektywność jest podobna. Brutlag i Meek[11] porównali klasyfikatory k-means, SVM oraz naive Bayes i stwierdzili, że różne zbiory danych powodują większą zmienność w dokładności klasyfikacji niż same algorytmy klasyfikacji.

Poprawy jakości klasyfikacji można oczekiwać od łącznego wykorzystania wielu niezależnych klasyfikatorów (*ensembles*)[12]. Są również prace, które niezgodność klasyfikatorów wykorzystują do ich automatycznego uczenia. Spamerzy zmieniają taktykę, dlatego też działanie klasyfikatorów pogarsza się z czasem. Standardową metodą walki ze spamem jest nauczenie klasyfikatora na podstawie ręcznie oznaczonych instancji, a to jest kosztowne. Chinavle i Kolari[13] proponują

---

6 M. Chang, C.K. Poon, *Using phrases as features in email classification*, „Journal of Systems and Software" 2009, vol. 82(6), s. 1036–1045, doi:10.1016/j.jss.2009.01.013.

7 S. Tong, D. Koller, *Support vector machine active learning with applications to text classification*, „The Journal of Machine Learning Research" 2002, vol. 2(1), s. 45–66, http://dl.acm.org/citation.cfm?id=94479.

8 K. Gee, *Using latent semantic indexing to filter spam*, w: *Proceedings of the 2003 ACM symposium on Applied computing*, 2003, s. 460–464, http://dl.acm.org/citation.cfm?id=952623.

9 F. Artigas-Fuentes, *Fast k-NN classifier for documents based on a graph structure*, „Progress in Pattern Recognition, Image Analysis, Computer Vision, and Aplication" 2010, vol. 6419, s. 228–235, http://www.springerlink.com/index/M3177401066H2337.pdf.

10 F. Sebastiani, *Machine learning in automated text categorization*, „ACM Computing Surveys" 2002, vol. 34(1), s. 1–47, doi:10.1145/505282.505283.

11 J.D. Brutlag, C. Meek, *Challenges of the Email Domain for Text Classification*, w: *ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning*, 2000, s. 103–110, http://dl.acm.org/citation.cfm?id=645529.657817.

12 R. Neumayer, *Clustering based ensemble classification for spam filtering*, 2006, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.140.9223&rep=rep1&type=pdf.

13 D. Chinavle, P. Kolari, *Ensembles in adversarial classification for spam*, w: *Proceedings of the 18th ACM conference on Information and knowledge management*, 2009, s. 2015–2018, http://dl.acm.org/citation.cfm?id=1646290.

właśnie wykorzystanie wielu klasyfikatorów do uczenia systemu. Zakładają, że niezgodność klasyfikatorów jest sygnałem, że pewna z cech spamu została wykluczona w wyniku zmiany taktyki spamerów.

O ile to możliwe, nie tylko słowa kluczowe z wiadomości powinny być wykorzystywane. Przede wszystkim zauważono, że spam e-mailowy i blogowy to różne rzeczy[14]. Ten drugi rodzaj spamu pozwolił na wykorzystanie dodatkowych cech: znaczników HTML, fragmentów URL, a do tego zbiorów słów, n-gramów słów, n-gramów znaków.

Aby dodać wpis na blogu, trzeba się zarejestrować. Można więc obserwować nie pojedyncze posty, lecz identyfikować ludzi je zamieszczających. W ramach *ECML/PKDD 2008 Challenge Discovery* jako wyzwanie postawiono identyfikację spamerów w systemie zakładek społecznych tak szybko, jak to możliwe. Opierając się na tym zbiorze, Kyriakopoulou i Kalamboukis[15] użyli SVM/TSVM do identyfikacji metacech poszczególnych podziałów, co było wykorzystane do klasyfikacji nadawców postów jako spamerów.

Najbardziej dojrzałym i zaawansowanym rozwiązaniem do filtrowania spamu jest SpamAssassin (SA)[16], przyjęty jako standard do ochrony serwerów poczty. Wykorzystuje różnorakie mechanizmy: od analizy zawartości treści oraz nagłówków wiadomości poprzez filtrowanie bayesowskie do analiz sieciowych. Używane są zarówno stałe reguły, jak i statystyki wyliczane na bieżąco ze strumienia wiadomości, co pozwala na doskonalenie systemu.

Do oceny, czy dana wiadomość jest spamem, SA wykorzystuje tzw. testy, których obecnie jest 711 (wersja 3.3)[17]. Każdy z testów ma przypisaną odpowiednią ocenę (*score*): dodatnią, jeśli spełnienie warunków danego testu zwiększa prawdopodobieństwo klasyfikacji jako spamu, i ujemną w przeciwnym przypadku. Wiadomość przechodzi wszystkie testy, a SA łączy oceny pozytywnych testów w ocenę globalną[18]. Skoring pojedynczego testu na spam najczęściej wynosi 1. Domyślna instalacja SA przyjmuje, że wiadomość jest klasyfikowana jako spam, jeśli ocena punktowa przekracza 5.

Poniższy przykład przedstawia nagłówki wiadomości, która została zaklasyfikowana jako poprawna wiadomość:

---

[14] Ibidem.

[15] A. Kyriakopoulou, T. Kalamboukis, *Combining clustering with classification for spam detection in social bookmarking systems*, 2008, http://ipl.cs.aueb.gr/publications/Combining Clustering with Classification for Spam Detection in Social Bookmarking Systems.pdf.

[16] http://spamassassin.apache.org/.

[17] Pełna lista reguł dostępna na: http://spamassassin.apache.org/tests_3_3_x.html.

[18] Nie zostało napisane wprost, jak wygląda łączenie, ale z dokumentacji wynika, że jest to prosta suma ocen cząstkowych.

X-Spam-Status: No, score=–2.5 required=5.0 tests=AWL,BAYES_00, HTML_80_90, HTML_MESSAGE,HTML_NONELEMENT_00_10 autolearn=ham version=3.0.2

Drugi przykład przedstawia nagłówki wiadomości, która została zaklasyfikowana jako spam. Dołączone zostało również wyjaśnienie oceny:

X-Spam-Status: Yes, score=5.3 required=5.0 tests=AWL,BAYES_95, HTML_10_20, HTML_MESSAGE,MIME_BOUND_NEXTPART,UNPARSEABLE_RELAY autolearn=no version=3.1.1
X-Spam-Report:
* 0.0 UNPARSEABLE_RELAY Informational: message has unparseable relay lines
* 0.0 HTML_MESSAGE BODY: HTML included in message
* 3.0 BAYES_95 BODY: Bayesian spam probability is 95 to 99% [score: 0.9606]
* 1.4 HTML_10_20 BODY: Message is 10% to 20% HTML
* 0.3 MIME_BOUND_NEXTPART Spam tool pattern in MIME boundary
* 0.7 AWL AWL: From: address is in the auto white-list

Oceny przypisane poszczególnym testom są wyliczane na podstawie wcześniej sklasyfikowanych wiadomości. Do uczenia wag w wersji SA 2.0 stosowane były algorytmy genetyczne, a w wersji 3.0 zastąpiono je siecią neuronową (perceptronem).

Ciągle toczy się dyskusja dotyczącego tego, czy reguły powinny być publiczne – spamerzy mogą to wykorzystać do obchodzenia mechanizmów. Jak się okazuje, nawet po opublikowaniu reguły niewiele programów do masowego wysyłania zmienia swoje działanie, co ciągle pozwala na poprawną klasyfikację spamu[19]. Dodatkowym atutem jest to, że reguły są wystawiane na wspólne ulepszanie w procesie *open source*. Sekretne reguły stosowane przez dostawców internetowych dużo częściej stają się celem ataku.

SpamAssassin jest doskonałym przykładem wykorzystania dodatkowej wiedzy, niepochodzącej z samej treści maila. Blokowanie spamu może odbywać się na podstawie zewnętrznych baz sygnatur, adresów IP (wykorzystując DNS), zabronionych adresów URL. Można uwiarygodnić też wiadomość poprzez wykorzystanie Sender Policy Framework, DomainKeys czy Hashcash. Dodatkowo test Bayesa daje skoring wiadomościom podobnym do tych, które wcześniej zostały uznane za spam.

Istotnym ograniczeniem przedstawionych wyżej metod jest brak uwzględnienia wiedzy dziedzinowej, którą dysponujemy. Opierają się one głównie na wykorzystaniu słów z treści (zbiór termów jest otwarty, indeksowanie

[19] http://taint.org/2005/08/06/024026a.html.

pełnotekstowe, następnie ważenie). Nie jest uwzględniana żadna struktura zewnętrzna (np. wspólny autor).

W projekcie SMC dość dobrze wiadomo, jakie wiadomości są poszukiwane: zbiór słów raczej zamknięty – nazwy leków są ściśle określone. To występowanie nazwy leku decyduje głównie o uznaniu czegoś za zagrożenie.

Rozważane było zastosowanie któregoś z podejść uczenia z nadzorem. Wymagałoby to oznaczenia każdego z ogłoszeń jako zagrożenie. Bez doświadczenia, jakie posiadają oficerowie śledczy, nie da się tego zrobić. Poza tym z rozmów wynika, że to, co jest napisane w warstwie leksykalnej, niesie coś innego w warstwie semantycznej. Ograniczenia tego żadna metoda nie może ominąć. Poza tym celem jest nie tyle klasyfikacja ogłoszeń, ile ich ranking według zagrożenia.

Dlatego, podobnie jak to jest w przypadku SpamAssassin, zdecydowano się na obiektywizację przesłanek, licząc się z pewnymi fałszywymi sygnałami. Minimalizacja ich liczby może być osobnym zagadnieniem badawczym. Wartością dodaną jest uwzględnienie słownika w ocenie oraz dodatkowa wiedza wynikająca z agregacji różnych ogłoszeń.

4. Jakościowa charakterystyka oceny zagrożenia

W projekcie SMC przyjęto, że wykryte zagrożenia będą oceniane jednowymiarową miarą służącą użytkownikom do porządkowania znalezionych przez system potencjalnych zagrożeń. Miara ta jest określana jako „**stopień ważności zagrożenia**" rozumiany jako istotność skutków niewykrycia i w konsekwencji doprowadzenia do materializacji zagrożenia. Oczywiście użytkownik ma do dyspozycji także inne metody filtrowania i sortowania zagrożeń, ale proponowana miara ma być ogólna, użyteczna w pierwszej fazie analizy dużej liczby nowych zgłoszeń. Jest istotna dla optymalizacji procesu szczegółowej analizy zagrożeń.

Elementy oceny wynikają z konstrukcji profilu zagrożenia. Jego trzy najważniejsze atrybuty to: podmiot ogłaszający, obiekt obrotu oraz akcja. Stopień ważności zagrożenia jest zatem określany na podstawie ocen ważności tych trzech atrybutów zagrożenia:

- **podmiot ogłaszający**: im częściej zamieszcza ogłoszenia, tym większą uwagę powinno się na nim skupić. Tożsamość podmiotu jest opisywana zbiorem wartości cech identyfikujących, np. e-mail, numer telefonu, identyfikator komunikatora internetowego. Podmiot ogłaszający jest rozumiany jako osoba fizyczna i jest w relacji jeden do wielu z nadawcą ogłoszenia. Ten drugi rozu-

miany jest jako użytkownik zarejestrowany pod określonym pseudonimem na danym forum lub portalu.

- **obiekt obrotu**: ze względu na brak ogólnie przyjętej klasyfikacji leków za wzmiankę o leku uznawane jest pojawienie się w ogłoszeniu dowolnej nazwy zdefiniowanej w systemowej tabeli nazw. Ze względu na występowanie w języku potocznym synonimów obiekty, które nie mają być rozróżnialne z punku widzenia oceny, powinny mieć zdefiniowaną wspólną „nazwę znormalizowaną". Na rozróżnianie leków nie wpływają dawki, postaci, opakowania itp. oraz oferowana w ogłoszeniu ilość leku. Na ocenę wpływa jedynie liczba różnych leków występujących w danym ogłoszeniu.
- **akcja**: rozpoznawane są oferty kupna lub sprzedaży. W ogłoszeniu może być rozpoznana więcej niż jedna akcja, a każda z nich może być powiązana z odrębną grupą obiektów. Ze względu na niską pewność rozpoznania takich związków (i przekonanie o małym znaczeniu takich powiązań dla użytkownika) ocenę zagrożenia wylicza się na podstawie najbardziej prawdopodobnego przypadku.

Wyliczanie oceny ogólnej zagrożenia wyłącznie na podstawie stopnia ważności zagrożenia wymagałoby założenia doskonałego rozpoznania intencji publikujących ogłoszenia przez system. Przesłanki, z których wynika ocena rozpoznania zebranych danych jako zagrożenia, są w znacznym stopniu niejednoznaczne i obarczone dużym prawdopodobieństwem błędu, dlatego proponuje się do opisu rozpoznawanych zmiennych (atrybutów zagrożenia) dołączać informację o „**pewności rozpoznania**" danej zmiennej[20].

5. Kwantyfikacja oceny zagrożenia oraz wpływu niepewności rozpoznania poszczególnych atrybutów profilu

W niniejszej sekcji dokonana jest kwantyfikacja stopnia zagrożenia. Składają się na nią trzy czynniki związane z opisanymi wcześniej atrybutami. Każdy z czynników przyjmuje wartość z zakresu <0; 1>.

W dalszej części surowe czynniki korygowane są o pewność rozpoznania. Poszczególne komponenty powinny umieć określić stopień pewności wygenerowanego wyniku, wyrażony wartością z przedziału <0; 1>. Na przykład

[20] Na przykład poprzez określenie prawdopodobieństwa poprawności działania danej reguły ekstrahującej wartość atrybutu.

rozpoznanie obiektu jako leku ma pewność 1,0, jeśli nazwa określająca obiekt występuje dosłownie w słowniku; 0,9 – jeśli w słowniku występuje nazwa zgodna z 90% liter nazwy z ogłoszenia.

Podmiot ogłaszający

Z założeń systemu wynika, że z ogłoszeniem związany jest tylko jeden nadawca. Z punktu widzenia pracy operacyjnej policji priorytetem jest ściganie osób stale trudniących się sprzedażą leków, a więc również zamieszczających wiele ogłoszeń. Chociaż nadawca nie ma wpływu na ocenę pojedynczego ogłoszenia, to jednak agregacja ogłoszeń może wnieść dodatkową informację. Istotny wpływ na skoring ma zatem liczba zagrożeń związanych z danym nadawcą. Przyjęto, że czynnik *S* związany z nadawcą wyrażony jest wzorem:

$$S = \begin{cases} 0,2n & \text{dla } n \le 5 \\ 1,0 & \text{dla } n > 5 \end{cases},$$

gdzie:
n – liczba postów o skoringu ≥0,5 występujących w bazie zagrożeń i przypisanych do danego nadawcy.

Osoby ogłaszające starają się nie ujawniać skali swojej działalności, dlatego za wszelką cenę starają się maskować swoją tożsamość, np. zamieszczając takie same posty z różnymi pseudonimami. Dobre wyliczenie czynnika *S* jest więc pośrednio związane z problemem tożsamości podmiotu ogłaszającego związanego z nadawcą. Jest to jednak zagadnienie badawcze wykraczające poza ramy tego artykułu, niemniej w systemie uwzględnione.

Niepewność rozpoznania nadawcy (a pośrednio tożsamości podmiotu ogłaszającego) powinna mieć najmniejszy wpływ na ocenę danego profilu jako zagrożenia. Obniżenie skoringu może mieć na celu jedynie wprowadzenie uporządkowania działań doraźnych, ale nie eliminację profilu. Dobre rozpoznanie atrybutu sprzyja jednoznacznej identyfikacji i może prowadzić do szybszych działań operacyjnych. Czynnik skorygowany o niepewność rozpoznania to:

$$S' = (0,8 + 0,2 \cdot c_s) \cdot S,$$

gdzie:
c_s – pewność rozpoznania tożsamości nadawcy; wartość z przedziału <0; 1>.

Taka postać wzoru oznacza, że współczynnik korygujący mieści się w przedziale <0,8; 1>. Jeśli pewność rozpoznania atrybutów nadawcy wynosi 0 (np. brak nadawcy), to wartość czynnika *S* obniżana jest do wartości 0,8**S*.

Obiekt obrotu

Duża liczba leków w pojedynczym ogłoszeniu może wskazywać na ciągły charakter danej działalności. Podobnie rzecz ma się z dużymi ilościami leków oferowanymi w pojedynczym ogłoszeniu lub wręcz określeniami „dowolna ilość". Dodatkowo dla oceny zagrożenia istotne są dwa typy obiektu: lek oraz recepta, którym przypisane jest waga *w* równa odpowiednio 1,0 oraz 0,5.

Dla zbioru obiektów znormalizowanych wykrytych w ogłoszeniu tworzy się ocenę wg wzoru:

$$O = \begin{cases} 0,2 \cdot w \cdot n & \text{dla } n \leq 5 \\ w & \text{dla } n > 5 \end{cases},$$

gdzie:

n – łączna liczba różnych znormalizowanych obiektów występujących w ogłoszeniu;

w – waga danego typu obiektu.

Niepewność rozpoznania typu obiektu powinna mieć podstawowy wpływ na uznanie danego profilu za zagrożenie. Czyli, jeśli pewność rozpoznania, że obiekt operacji należy do typów istotnych wynosi 0, to skoring również powinien wynieść 0. Stąd proponowany przedział wartości dla współczynnika korygującego <0; 1>.

Czynnik skorygowany o niepewność rozpoznania to:

$$O' = c_o \cdot O,$$

gdzie:

c_o – pewność rozpoznania obiektu jako leku; wartość z przedziału <0; 1>.

Akcja

W ustawie brakuje sankcji za kupno leku. Teoretycznie więc ogłoszenia kupna można by pomijać. Z wywiadów z oficerami operacyjnymi wynika jednak, że czasami „kupię" oznacza tak naprawdę „sprzedam". Szczególnie podejrzane mogą być oferty kupna, przy których pojawia się cena. Prawdziwe intencje mogą być dopiero określane przez ludzi. Zdecydowano o uwzględnianiu ofert kupna, choć z mniejszą wagą.

Z jednym zagrożeniem może być związanych od 0 do *n* akcji, odpowiadających poszczególnym frazom (warstwa leksykalna). Pierwotnie w systemie starano się wiązać każdy obiekt z akcją, na poziomie leksykalnym, np. „sprzedam W, oferuję też Z". Znajdowano ogłoszenia, w których ktoś jednocześnie sprzedawał

obiekt X i chciał kupić obiekt Y (zamiana). Ze względu na trudności z rozpoznawaniem poszczególnych związków wszystkie rozpoznane obiekty są oceniane łącznie, a do wyliczania oceny wybierany jest najbardziej prawdopodobny typ akcji (warstwa semantyczna). Pod uwagę brane są wartości po normalizacji, tj. usunięciu literówek, uwzględnieniu synonimów.

Proponuje się przyjęcie następujących wartości dla czynnika związanego z akcją:

$$A = \begin{cases} 1{,}0 & \text{gdy akcja = sprzedaż} \\ 0{,}75 & \text{dla nierozpoznanej akcji} \\ 0{,}5 & \text{gdy akcja = kupno} \end{cases}.$$

Niepewność rozpoznania akcji jako sprzedaży lub kupna powinna mieć mniejszy wpływ na uznanie danego profilu za zagrożenie, niż jest to w przypadku obiektu operacji. Wynika to z dużego prawdopodobieństwa maskowania rzeczywistej operacji (użycie slangu) i większej różnorodności językowej dla wyrażenia tego pojęcia.

Czynnik skorygowany o niepewność rozpoznania to:

$$A' = (0{,}5 + 0{,}5 \cdot c_a) \cdot A,$$

gdzie:

c_a – pewność rozpoznania akcji; wartość z przedziału <0; 1>.

Taka postać wzoru oznacza, że współczynnik korygujący mieści się w przedziale <0,5; 1>. Jeśli pewność rozpoznania akcji wynosi 0 (np. brak akcji), to wartość czynnika A obniżana jest do wartości 0,5*A.

Ocena łączna

Wstępnie jako ocenę łączną zaproponowano kombinację liniową poszczególnych czynników:

$$SC(x) = \alpha O' + \beta S' + \gamma A'$$
$$\alpha + \beta + \gamma = 1$$
$$\alpha, \beta, \gamma \geq 0,$$

gdzie:

x – profil zagrożenia,

α, β, γ – wagi przypisane poszczególnym czynnikom zagrożenia,

S', O', A' – skorygowane czynniki zagrożenia.

Parametry służą lepszemu dopasowaniu skoringu do wymagań użytkowników. Użytkownicy będą mieli możliwość parametryzacji (modyfikowania wartości parametrów) wzoru na wyznaczenie oceny zagrożenia, ale nie samodzielnego definiowania takiej oceny.

Wadą rozwiązania liniowego jest to, że ocena łączna nie zeruje się, jeśli któryś z czynników ma ocenę 0. W związku z tym zaproponowano ważony iloczyn:

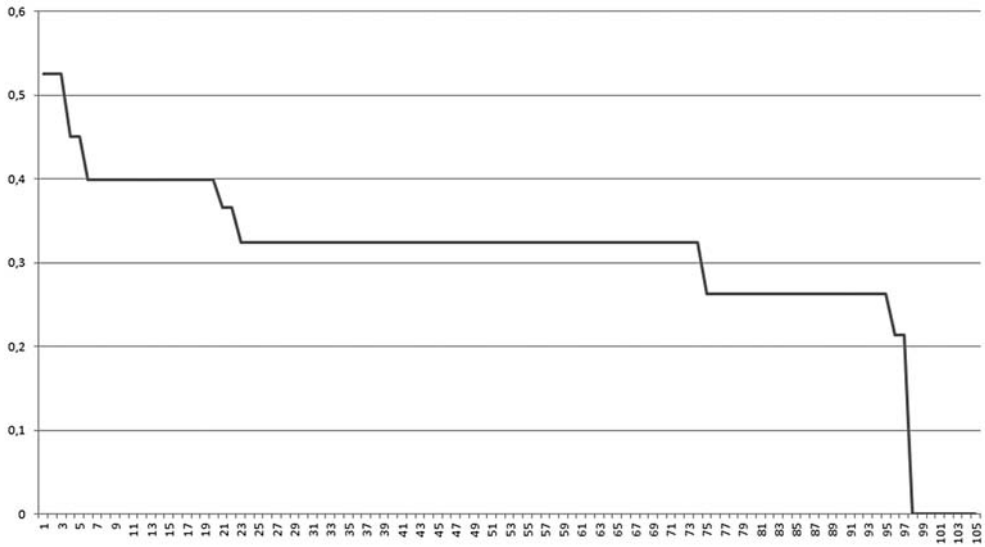
$$SC(x) = O'^{\alpha} \cdot S'^{\beta} \cdot A'^{\gamma}$$
$$\alpha + \beta + \gamma = 1$$
$$\alpha, \beta, \gamma \geq 0.$$

Jeśli w ogłoszeniu nie występują informacje o lekach ($O = O' = 0$), to również skoring takiego ogłoszenia wynosi 0 i nie jest ono klasyfikowane jako zagrożenie.

6. Implementacja i wyniki

Implementacja przedstawionych wyżej formuł została bezpośrednio podporządkowana wymogom wydajnościowym systemu – strumień zagrożeń może być znaczny. Do przechowywania danych w systemie została wybrana baza PostgreSQL. Pierwotny pomysł zakładał napisanie aplikacji w Javie łączącej się z bazą za pomocą JDBC. Ze względu na to, że w obliczeniach są głównie wykorzystywane operacje arytmetyczne oraz operacje na bazie danych, ostatecznie miejsce obliczeń zostało umiejscowione jak najbliżej bazy. Została zaimplementowana odpowiednia funkcja bezpośrednio w bazie w języku PL/pgSQL. Daje to tę dodatkową korzyść, że można wykorzystać mechanizmy bazodanowe, takie jak harmonogram czy zastosowanie wyzwalaczy do uruchamiania obliczeń skoringowych. Skoring zależy głównie od zawartości danego ogłoszenia, a zmiany obejmujące wiele rekordów zachodzą wtedy, gdy pojawia się post od znanego już wcześniej autora. Niemniej ze względu na strukturę bazy można określić, które rekordy należy zaktualizować, do czego została zaproponowana odpowiednia metoda. Implementacja w PL/pgSQL jest więc najwydajniejszym rozwiązaniem, zapewniającym odpowiednią skalowalność systemu.

Poniższy rysunek przedstawia rozkład wag dla pewnego losowo wybranego zestawu ogłoszeń. Ze względu na przyjętą formułę może się zdarzyć, że wiele postów ma identyczną wagę, co utrudnia nieco ich priorytetyzację. Wprowadzenie dodatkowych kryteriów różnicujących może być jednym z celów dalszych badań.



Rysunek 3. Rozkład ocen zagrożeń dla losowo wybranych profili
Źródło: opracowanie własne.

7. Podsumowanie i kierunki dalszych badań

W niniejszym artykule zaproponowano metodę oceny zagrożenia związanego z publikacją ogłoszeń. Kierowano się przede wszystkim tym, aby ocena ogólna była prosta, a więc jednowymiarowa. Z tego powodu reprezentuje cechy istotne dla wszystkich wykrytych wystąpień zagrożenia.

Można oczywiście dyskutować z wieloma przyjętymi tutaj założeniami. Przyjęto jednak podejście pragmatyczne, opierające się na doświadczeniu osób na co dzień zajmujących się ściganiem przestępstw, których ślady pozostawiane są na forach internetowych. Przyjętą ocenę należy traktować bardziej w kategoriach względnych – ma pomagać w szeregowaniu ogłoszeń do sprawdzenia. Arrow już w 1951 r. wykazał, że nie ma idealnego rankingu. Zdefiniował cztery kryteria dobrej oceny i okazało się, że wszystkie nie mogą być jednocześnie spełnione[21].

Osobnym problemem badawczym jest określenie tożsamości nadawcy, tj. przypisanie go do podmiotu ogłaszającego. Ma to zasadnicze znaczenie dla

[21] K.J. Arrow, *Social Choice and Individual Values*, Yale University Press, 1951.

oceny zagrożenia. Przyjęto roboczo, że zgodna tożsamość nadawcy jest określona przez: numer telefonu (1 lub więcej), e-mail (1 lub więcej), komunikator (1 lub więcej), podobieństwo tworzonych postów powyżej określonego progu lub przez decyzję użytkownika.

Niewątpliwie lepszym sposobem na stworzenie rankingu jest bezpośrednia konfrontacja, czyli porównanie dwóch ogłoszeń i zdecydowanie, które stanowi większe zagrożenie. Tutaj w ocenie niezbędni są ludzie: lepiej rozpoznają rodzaj leku, dawkę, cenę czy wreszcie próbę maskowania właściwej treści ogłoszenia. Zaplanowane są badania, które pozwolą na wykorzystanie metody Elo[22] do utworzenia rankingu ogłoszeń. Użytkownicy mogą ocenić, czy sprzedawane obiekty stanowią zagrożenie. Nie są jednak w stanie uwzględnić wiedzy o nadawcach – nie zapamiętają różnych numerów telefonów czy też podobnych szablonów ogłoszeń. W badaniu tym zostaną zatem wstępnie zebrane informacje o nadawcach i będą prezentowane użytkownikom łącznie z samą treścią ogłoszenia. Połączenie dwóch podejść może dać ciekawe wyniki.

Bibliografia

1. Arrow K.J., *Social Choice and Individual Values*, Yale University Press, 1951.
2. Artigas-Fuentes F., *Fast k-NN classifier for documents based on a graph structure*, „Progress in Pattern Recognition, Image Analysis, Computer Vision, and Aplication" 2010, vol. 6419, s. 228–235, http://www.springerlink.com/index/M3177401066H2337.pdf.
3. Brutlag J.D., Meek C., *Challenges of the Email Domain for Text Classification*, w: *ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning*, 2000, s. 103–110, http://dl.acm.org/citation.cfm?id=645529.657817.
4. Chang M., Poon C.K., *Using phrases as features in email classification*, „Journal of Systems and Software" 2009, vol. 82(6), s. 1036–1045, doi:10.1016/j.jss.2009.01.013.
5. Chinavle D., Kolari P., *Ensembles in adversarial classification for spam*, w: *Proceedings of the 18th ACM conference on Information and knowledge management*, 2009, s. 2015–2018, http://dl.acm.org/citation.cfm?id=1646290.
6. Elo A., *The Rating of Chessplayers, Past and Present*, Arco 1978.
7. Gee K., *Using latent semantic indexing to filter spam*, w: *Proceedings of the 2003 ACM symposium on Applied computing*, 2003, s. 460–464, http://dl.acm.org/citation.cfm?id=952623.

[22] A. Elo, *The Rating of Chessplayers, Past and Present*, Arco 1978.

8. Karras D., *An improved text categorization methodology based on second and third order probabilistic feature extraction and neural network classifiers*, „Knowledge-Based Intelligent Information and Engineering System" 2006, vol. 4251, s. 9–20, http://www.springerlink.com/index/07m8415wj15v2677.pdf.
9. Kyriakopoulou A., Kalamboukis T., *Combining clustering with classification for spam detection in social bookmarking systems*, 2008, http://ipl.cs.aueb.gr/publications/Combining Clustering with Classification for Spam Detection in Social Bookmarking Systems.pdf.
10. Małyszko J., Filipowska A., Abramowicz W., Kaczmarek T., Bukowska E., Perkowski B., Stolarski P. et al., *Architektura systemu wykrywania zagrożeń w cyberprzestrzeni*, „Roczniki" Kolegium Analiz Ekonomicznych SGH, z. 24, Oficyna Wydawnicza SGH, Warszawa 2012, s. 11–22.
11. Markov A., *Fast categorization of Web documents represented by graphs*, „Advances in Web Mining and Web Usage Analysis" 2007, vol. 4811, s. 56–71, http://www.springerlink.com/index/u4886005r4760437.pdf.
12. Neumayer R., *Clustering based ensemble classification for spam filtering*, 2006, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.140.9223&rep=rep1&type=pdf.
13. Ogura H., Amano H., Kondo M., *Feature selection with a measure of deviations from Poisson in text categorization*, „Expert Systems with Applications" 2009, vol. 36(3), s. 6826–6832, doi:10.1016/j.eswa.2008.08.006.
14. Sebastiani F., *Machine learning in automated text categorization*, „ACM Computing Surveys" 2002, vol. 34(1), s. 1–47, doi:10.1145/505282.505283.
15. Senator T.E., *On the efficacy of data mining for security applications*, w: *Proceedings of the ACM SIGKDD Workshop on CyberSecurity and Intelligence Informatics – CSI-KDD '09*, ACM Press, New York 2009, s. 75–83, http://dl.acm.org/citation.cfm?id=1599272.1599286.
16. Tong S., Koller D., *Support vector machine active learning with applications to text classification*, „The Journal of Machine Learning Research" 2002, vol. 2(1), s. 45–66, http://dl.acm.org/citation.cfm?id=944793.

Źródła sieciowe

1. http://spamassassin.apache.org [dostęp 07.08.2012].
2. *The life of Spam Assassin Rule*, http://taint.org/2005/08/06/024026a.html [dostęp 08.08.2012].

* * *

The method for scoring of threat related to online adverts publication

Summary

The paper proposes a method for scoring adverts available in online portals as a threat understood as the possibility of breaking the law. We specifically focus on illegal drug trade scenario. The combined score bases on three components scored separately: person publishing the offer, traded good, and intended action (buy, sell, other).

Keywords: scoring, threat, advert, cyberspace monitoring