Klasyfikator Stron Internetowych jako system informatyczny ograniczający dostęp do internetu przy wykorzystaniu klasyfikatora słów kluczowych z wagą

Piotr Wójcik
PJA,
Warszawa, piotr.wojcik@pja.edu.pl

8 lipca 2018

1 Abstrakt

Artykuł opisuje nową metodę przeszukiwania stron Internetowych w poszukiwaniu stron niebezpiecznych dla Internauty. Zastosowano tu słowa kluczowe z parametrem wagi w celu klasyfikacji a nawet cenzury dostępu do Internetu dzięki liście stron zablokowanych umieszczonych na przygotowanych serwerach DNS.

Słowa kluczowe: *Internet, klasyfikator, klasyfikacja, słowa kluczowe z wagą, DNS*

2 Wstęp

Niemal każdą zaindeksowaną stronę w Internecie można znaleść w odpowiedniej wyszukiwarce za pomocą **słów kluczowych** (ang. **keywords**). Z tego powodu istnieje możliwość sklasyfikowania strony internetowej za pomocą odpowiednich słów. Stąd stosunkowo jest łatwo odkreślić dla jakiego użytkownika Internetu przeznaczona jest dana strona internetowa.

W tym artykule chciałbym przybliżyć podstawy rozpoznawania stron Internetowych za pomocą **słów kluczowych z wagą**. Na podstawie analizy występowania słów kluczowych zwanych także **frazą wyszukiwania** (ang. **search phrase**) można określić dla jakiej kategorii wiekowej dopuszczalna jest strona. Można to określić jawnie bądź przez domniemanie. Na przykład jeśli na stronie występuje słowo„nazizm" to bardzo prawdopodobnie strona ta nie będzie odpowiednia dla użytkownika w wieku do 14 lat. Jest to jawny warunek wykluczający. Klasyfikacja za pomocą domniemania jest bardziej skomplikowana. Jeśli natomiast w stronie Internetowej występuje fraza „Adolf Hitler" to prawdopodobnie przez skojarzenie strona będzie zakwalifikowana jako strona o nazizmie i stąd także będzie podlegała odpowiedniej klasyfikacji. Jednak powyższa fraza może dotyczyć strony z biografią Adolfa Hitlera i stąd może być odpowiednia dla uczniów szkoły średniej jednak wciąż fraza jest ona na tyle silna by wykluczyć młodego Internautę.

Rodzi się tu problem odpowiedniej klasyfikacji występujących słów kluczowych. Weźmy na przykład słowo „piersi", które może odnosić się do stron internetowych o na przykład tematyce erotycznej lub wręcz pornograficznej, ale nie możemy tego jednoznacznie stwierdzić, ponieważ może ono wystąpić w sąsiedztwie słów „rak piersi" i być stroną o tematyce medycznej. Jest to przykład możliwej klasyfikacji nie aż tak silnej jak w przypadku stron ze słowem kluczowym „Hitler", ponieważ w rozróżnieniu od stron pornograficznej strony o tematyce medycznej mogą być dopuszczone dla Internauty w wieku licealnym a być może i młodszym. Dlatego istnieje potrzeba odpowiedniego pogrupowania słów kluczowych. Mogą nam pomóc w tym odpowiednie słowniki wraz z określeniem wag, które będą rozróżniały słowa kluczowe klasyfikujące silnie od fraz słabszych, które mogą jedynie naprowadzić na tematykę strony internetowej.

Można też przeprowadzić klasyfikację stron za pomocą sztucznych sieci neuronowych, które dla przykładu rozpoznają nagość na zdjęciach i filmach [1–3]. Technika wykrywania pokrycia ciała skórą w rozpoznawaniu nagości cieszy się popularnością i zdaje dobre wyniki [4]. Ta tematyka nie będzie poruszana w tym artykule.

Co prawda wszelka cenzura w Internecie nie się za sobą kontrowersję [5]. ale co raz więcej nie reżimowych państw korzysta z tej techniki jak Stany zjednoczone, Wielka Brytania i Australia. Jednym z powodów dlaczego państwa demokratyczne skłaniają się do mniejszego zła jakim jest ograniczanie dostępu do Internetu dla osób najmłodszych jest ilość niebezpieczeństw jakie może on tam znaleźć. [6]

3 DNS

Najlepszą i za razem najbezpieczniejszą formą ograniczenia dostępu do odkrywania stron Internetowych jest zastosowanie odpowiednio przygotowanych serwerów Domain Name System DNS. [7–9]

Oczywiście bardzo łatwo można obejść to zabezpieczenie wpisując bezpośrednio w przeglądarce numer strony w postaci adresu IP, ale trzeba wcześniej ten numer znać. Jeśli chodzi o sam numer to w adresie IP wersji czwartej mamy do czynienia z 256^4 ilością możliwych domen a przy adresie IP wersji 6 ta pojemność wzrasta do niewyobrażalnej liczby $3.4 * 10^{1038}$. W praktyce liczba możliwych do zaadresowania stron internetowych jest niższa z powodu potrzeby uwzględnienia adresów różnych klas podsieci, które uniemożliwiają użycie niektórych adresów.

Drugim sposobem jest zamiana serwera DNS w chronionym komputerze, ale i ta operacja wymaga często uprawnień administratora.

4 Słowa kluczowe z wagą

Słowa kluczowe z wagą oprócz swojego znaczenia zapisanego w postaci słownej można rozszerzyć o krotkę posiadającą miarę istotności danego słowa w rozpatrywanym aktualnie algorytmie. Dla słowa „sex" możemy nadać na przykład wartoć 0.6 tworząc krotkę („sex", 0.6), która będzie wartością dla obiektu klasy WagedKeyword. Niech klasa ta zbudowana będzie z krotek, których pierwszym elementem będzie słowo w postaci ciągu znaków i wartością p taką, że $p \in \langle 0, 1 \rangle$. Im większy jest podany parametr p tym bardziej dane słowo jest brane pod

uwagę podczas ustalenia iloci punktów danej strony i większe prawdopodobieństwo do zakfalifikowania danej strony jako niebezpieczną.

5 Ustalanie wag oraz wydobywanie nowych słów kluczowych

Cechą języka naturalnego jest, że wciąż się rozwija. Tworzone są nowe słowa lub idiomy. Część słów wychodzi z obiegu stając się anachronizmami. Dlatego bardzo ważne jest aby słownik słów kluczowych i jego wagi były stale uaktualniane. W bazie danych klasyfikację stron Internetowych zaczynamy z zdefiniowanym słownikiem początkowym. Na jego podstawie możemy sklasyfikować część stron internetowych na podstawie, których możemy prowadzić dalszą ekstrakcję słów kluczowych. Cechą podobnych stron internetowych jest fakt, że używają podobnych słów kluczowych. Na tej podstawie można za pomocą odpowiedniego algorytmu ustalić aktualne wagi dla fraz wyszukiwania. O ile ustalenie wag jest czynnością stosunkowo prostą to już o wiele trudniejszym zadaniem jest wydobycie nowych fraz dla słownika klasyfikującego. Wymaga to zbudowania słownika słów oraz fraz występujących na każdej stronie wraz z ich liczebnością i porównać ten słownik z podobnymi stronami. Nie jest to zadanie trywialne i wymaga znacznych zasobów sprzętowych.

Do rozważenia przedstawiam szkic algorytmu, który wymaga dalszej dyskusji i ulepszeń.[alg1]

Alg 1. Podstawowy algorytm klasyfikacji stron internetowych

```
1
2  //input:
3  //     phrase - analysed phrase,
4  //     analysedPage - web page with URL and other characteristics,
5  //     compromisedWebPages - colection of compromised web pages,
6  //     redFlagPhraseDictionary - colection of phrases
7  //                               that compromise web page
8  //output:
9  //     score - points of page classification
10 //     compromisedWebPages - modyfied input object
11
12 mostCompromisedPages = compromisedWebPages.top();
13 foreach (phrase in webPageTextContent)
14 {
15   if((phrase in redFlagPhraseDictionary)
16       || (phrase like redFlagPhraseDictionary))
17   {
18     redFlagPhraseDictionary.computeWage(phrase, mostCompromisedPages);
19     score += redFlagPhraseDictionary.wage(phrase);
20     compromisedWebPages += analysedPage;
21   }
22 }
```

Algroytm ten jest zapisany w języku podobnym do C# z jedną różnicą. Brak w nim operatora like jaki możemy spotkać w językach 4 generacji jak na

przykład SQL, ale możliwe jest jego zastąpienieza pomocą metody like(). Zaimplementować ją można na wiele różnych sposobów. Na przykład korzystając z algorytmu Soundex stworzonego przez Roberta Russella i Margaret Odell [10]. Algorytm ten zwraca podobieństwo fonetyczne wyrazów w postaci 4-znakowego kodu okrelający odległość fonetyczne podobieństwo tych wyrazów.

Idea algorytmu polega na przeanalizowaniu wszystkich fraz *phrase* analizowanej strony internetowej *analysedPage* pod kątem występowania lub podobieństwa z zbiorem słów będących czerwoną flagą *redFlagPhraseDictionary* (linia 15). Gdy fraza jest podobna lub w szczególności ma taką samą wartość alfanumeryczną to:

1. aktualizujemy wagi naszego klasyfikatora w klasie *RedFlagPhraseDictionary* i aktualizujemy ranking w obiekcie klasy *MostCompromisedPages* (linia 18),
2. doliczamy punkty*score* zdobyte podczas analizowania strony (linia19),
3. dołączamy stronę do stron skompromitowanych (linia 20),

Linia 12 pobiera ranking najbardziej skompromitowanych stron *compromisedWebPages.top()* i jej słów *phrase*, które będą brane pod uwagę podczas każdej iteracji pętli algorytmu.

6 Kategorie stron internetowych (grupy stron internetowych)

Strony internetowe możemy podzielić na następujące grupy stron:

- strony informacyjne,
- internetowe bazy danych,
- strony komercyjne (sklepy),
- serwisy społecznościowe,
- portale randkowe,
- strony erotyczne,
- strony pornograficzne,
- strony z ryzykiem (na przykład rózne odmiany kasyn internetowych),
- strony z mową nienawiści,
- strony z nielegalną treścią,
- strony z nienawiścią oraz wszelką przemocą.

Nie jest to finalna lista i prawdopodobnie można zaproponować bardziej złożoną klasyfikację stron internetowych. Każdy algorytm klasyfikujący będzie używał innego zestawu wzorcowych słów klasyfikujących wraz z odpowiednimi wagami po to aby w przypadku zakwalifikowania strony jako niebezpiecznej dla Internauty można było opisać ją i analizować pod kątem odpowiedniego algorytmu.

7 Moduły Systemu Informatycznego Klasyfikatora Stron Internetowych

Klasyfikator Stron Internetowych w skrócie KSI (ang. Web Page Classification Information System WPCIS) będzie składać się kilku baz danych i programach wykorzystujących algorytmy potrzebne do przetwarzania danych.

Pierwszym modułem będzie baza danych stron nieprzetworzonych lub wstępnie przetworzonych. Zawierać będzie ona graf reprezentujący odkrytą część Internetu wraz z jej odpowiednim kolorem. Kolor określać będzie stronę nieprzejrzaną, przejrzaną wstępnie przez jeden z początkujących algorytmów, stronę w pełni sklasyfikowaną i stronę przetworzoną wraz z jej dalszymi odnośnikami. Strona sklasyfikowana należała będzie do odpowiedniej kategorii wraz z uzasadnieniem, które określi na podstawie, których najważniejszych słów kluczowych została oceniona w raz z ich liczebnością.

Drugi moduł będzie posiadał kluczowe znaczenie ze względu na przechowywanie słownika klasyfikujących słów kluczowych wraz z ich wagami. Baza ta w toku działania algorytmów będzie bardzo często uaktualniania oraz będzie stanowiła serce dynamicznego systemu ocen KSI. Bardzo ważne tu będzie przechowywanie perspektyw, które będą opisywały datę aktualnego algorytmu klasyfikującymi wraz z używanymi wagami oraz informację do jakiej kategorii klasyfikuje frazę.

Trzeci moduł to moduł określający wagi słów kluczowych. Jego zadaniem będzie ciągła aktualizacja wag słów kluczowych na podstawie przetworzonych słów z sklasyfikowanych stron internetowych.

Czwarty moduł to baza danych, na której głównie będą dopisywane już przetworzone strony wraz z ich klasyfikacją oraz uzasadnieniem. W pewnych przypadkach klasyfikacja strony może ulec zmianie stąd istnieje potrzeba zapewnienia tej strony ciągłej możliwości aktualizacji danych. Jest to bardzo ważny element systemu, ponieważ jest to moduł na podstawie, którego będzie budowany serwer DNS systemu.

Piąty oraz najtrudniejszy do zaimplementowania jest moduł prowadzący ekstrakcję nowych słów kluczowych.

Oczywiście wraz z rozrastaniem się wspomnianego systemu informatycznego dane moduły będzie można podzielić na drobniejsze. Na przykład najczęściej wykorzystywany i wymagający niezawodności moduł czwarty można podzielić na bazy danych stron internetowych należących do odpowiedniej kategorii (grupy stron). Na tej podstawie można zbudować bazę danych pod system DNS, dla odpowiedniej grupy wiekowej lub wykluczających dostęp do odpowiedniej popularnej grupy stron internetowych.

8 Zastosowanie Klasyfikatora Stron Internetowych

Pierwszym oczywistym miejscem zastosowania klasyfikatora będzie środowisko domowe, gdzie potrzeba zastosowania bezpiecznego Internetu ze względu na posiadanie małych dzieci jest niezbędnie potrzebna. W drugiej kolejności miejsce gdzie ów klasyfikator spełni swoją rolę jest środowisko pracy lub wszelkie placówki edukacyjne, gdzie pracodawca prawdopodobnie by chciał ograniczyć pracownikom dostęp do niektórych kategorii stron internetowych.

9 Podsumowanie i plany na przyszłość

Celem tego artykułu jest zaproponowanie nowej metody klasyfikacji stron internetowych. Planem autora jest zbudowanie działającego systemu bazującego na tym artykule. Najważniejszym produktem tego systemu będą gotowe serwery DNS posiadające wpisy do odpowiednich, dopuszczony grup stron internetowych. Umożliwi to takie skonstruowanie serwerów DNS, dzięki którym komunikacja do stron z nieodpowiednią treścią będzie niemożliwa lub znacznie utrudniona.

10 Bibliografia

[1] Will Archer Arentz , Bjørn Olstad. Classifying offensive sites based on image content.

[2] Radhouane Guermazi , Mohamed Hammami, Abdelmajid Ben Hamadou. Combining classifiers for web violent content detection and filtering.

[3] Giuseppe Amato , Pablo Bolettieri , Gabriele Costa , Francesco la Torre , Fabio Martinelli. Detection of images with adult content for parental control on mobile devices.

[4] Mohammad Reza Mahmoodi. High performance novel skin segmentation algorithm for images with complex background.

[5] Jonathan Zittrain, Benjamin Edelman. Internet filtering in china.

[6] Piotr Łuczuk. *Cyberwojna*.

[7] J. Postel, J. Reynolds. Domain requirements.

[8] P. Mockapetris. Domain Names - Concepts and Facilities.

[9] P. Mockapetris. Domain names - implementation and specification.

[10] Donald E. Knuth. *The Art of Computer Programming*, wolumen 3.