

# Wyszukiwanie i Przetwarzanie Informacji WWW

## Społeczne i ekonomiczne aspekty wyszukiwarek (2): Spam wyszukiwarkowy

Marcin Sydow

PJWSTK

# Plan tego wykładu

- Przyczyny istnienia spamu w WWW
- Techniki spammerskie
- Przykłady
- Walka ze spamem

# Mechanizmy WWW - przypomnienie

Przeprowadźmy następujący tok rozumowania, dotyczący stron komercyjnych:

- Strony WWW są odnajdywane za pomocą **wyszukiwarek**
- **widoczność** strony w wyszukiwarce wpływa na **wielkość ruchu** na stronie
- **wielkość ruchu** (ang. traffic) na stronie **przekłada się na zysk**

# Spam - motywacja

**Widoczność** strony w wyszukiwarkach zależy m.in. od następujących czynników:

- dla jakich zapytań dana strona zwracana jest jako wynik zapytania w wyszukiwarkach
- **jak wysoko w rankingach** dana strona pojawia się w wynikach wyszukiwania

Stąd istnieje silna motywacja do takiej modyfikacji stron aby pojawiały się jako wyniki określonych zapytań i żeby ich ranking był jak najlepszy.

# Spam - motywacja

**Widoczność** strony w wyszukiwarkach zależy m.in. od następujących czynników:

- dla jakich zapytań dana strona zwracana jest jako wynik zapytania w wyszukiwarkach
- **jak wysoko w rankingach** dana strona pojawia się w wynikach wyszukiwania

Stąd istnieje silna motywacja do takiej modyfikacji stron aby pojawiały się jako wyniki określonych zapytań i żeby ich ranking był jak najlepszy.

**Wartość tej “motywacji” określana jest na:**

**4.5 milarda dolarów** (Google, 2004)

## Co to jest Spam?

Celem jest “oszukanie” algorytmów rankingowych wyszukiwarki (niewykryte przez wyszukiwarke).

Przez **Spam Wyszukiwarkowy** (ang. Search Engine Spam) rozumie się:

**Celowe modyfikacje dokumentów WWW mające na celu “sztuczne” poprawienie pozycji rankingowej w wyszukiwarkach bez poprawy jakości informacyjnej z punktu widzenia zwykłego użytkownika**

Jest to nieprecyzyjne pojęcie. Inne określenie:

**“Wszystko co nie byłoby robione, gdyby nie istniały wyszukiwarki”**

# Spam a pozycjonowanie

Nie wszystkie techniki poprawy pozycji danej strony w wynikach wyszukiwania uznawane są za spam.

Wszelkie techniki mające na celu “legalną” poprawę jakości strony, lub ułatwienie dokładniejszego obliczenia stopnia dopasowania danej strony do określonych zapytań przez algorytmy rankingowe są nazywane (**legalnym**) **pozycjonowaniem** (ang. SEO - search engine optimization), o ile są zgodne z wytycznymi publikowanymi przez daną wyszukiwarkę.

Należy jednak pamiętać, że istnieje tylko “cienka czerwona linia” pomiędzy tym co uważane jest za legalne pozycjonowanie a tym co jest spamem. Należy zachować umiar i rozsądek.

# Spam a wyszukiwarki

Spam stanowi poważny problem dla wyszukiwarek, gdyż obniża jakość ich usług - w efekcie **obniża ich zyski**

Ze strony wyszukiwarek podejmowane są zdecydowane działania mające na celu **walkę ze spamem** (ang. search engine spam combating)

Można wymienić rozmaite aspekty tej walki:

- publikowanie wytycznych dla twórców stron
- wykrywanie spamu - rozwijanie i stosowanie automatycznych, “inteligentnych” technik
- zapewnianie mechanizmów zgłaszania spamu
- proponowanie pewnych standardów ułatwiających walkę ze spamem (np. “no follow”)
- reagowanie na wykryty spam



## Reagowanie na spam

Reakcją na wykryte przez wyszukiwarkę dokumenty spammerskie jest na ogół **pełne usunięcie ich z indeksu** (tak że **nie będą one już wogóle widoczne w wynikach zapytań**)

Nieco łagodniejszą formą jest obniżanie rankingów, ale jest to rozwiązanie droższe dla wyszukiwarek i raczej rzadko stosowane.

W niektórych przypadkach konsekwencje wyciągane są nie tylko wobec dokumentu spammerskiego, ale **także wobec jego sąsiedztwa** (w grafie linków WWW).

Powoduje to czasami gorące dyskusje na temat zasadności takich działań.

## “Wyścig Zbrojeń”

Z jednej strony, spamerzy stosują coraz bardziej wyrafinowane (trudniejsze do wykrycia) techniki niedozwolonej manipulacji.

Z drugiej strony, wyszukiwarki stosują coraz bardziej wyrafinowane algorytmy rankingowe oraz techniki wykrywania i przeciwdziałania spamowi.

Obie “strony” konfliktu stopniowo uczą się od siebie nawzajem. Ważne jest utrzymywanie najnowszych technik **w ścisłej tajemnicy** (szczególnie od strony wyszukiwarek)

Sytuacja przypomina “wyścig zbrojeń”, choć należy pamiętać, że strony nie są tutaj symetryczne (**wyścig ten może nie mieć końca**).

Kres może położyć **dobry mechanizm oparty na “równowadze ekonomicznej”**

# Taksonomia

Ogólnie, techniki spammerskie można podzielić na 2 główne grupy:

- “podbijanie” (ang. boosting) - bezpośrednio zwiększanie szansy, że strona ukaże się wyżej w rankingach (wymiary: np. tekst i struktura linków)
- maskowanie (ang. hiding) - utrudnianie systemom anti-spamowym wyszukiwarek wykrycia zabronionych technik (tekstowe, linkowe, związane z protokołem HTTP: np. “podmiana” (ang. cloaking) czy przekierowania)

# Spamowanie Tekstu

Obserwuje się spamowanie wszystkich “kontekstów” tekstu w dokumentach WWW: ciało dokumentu (ang. body), tytuł, znaczniki meta, tekst odnośników (ang. anchor text), nazwa URL

Przykładowe techniki “podbijające” i maskujące:

- powtarzanie (cel: np. zwiększanie TF)
- zaśmiecanie - umieszczanie dużej ilości niezwiązanych terminów (cel: sprawianie, że dokument “pasuje” do bardzo wielu zapytań o rzadkich słowach kluczowych)
- “rozpuszczanie”(ang. dilution) spamowanych słów kluczowych w innym, “normalnym” tekście - utrudnia wykrywanie spamu
- stosowanie bardzo małych czcionek
- tekst w kolorze tła

# Spamowanie Linków

Można wyróżnić 3 grupy dokumentów (ze względu na dostępność dla spamera):

- własne (np. będące celem manipulacji)
- pozostałe dostępne (np. blogi, fora, etc.)
- niedostępne (pozostałe strony)

Techniki maskujące związane z linkami:

- ukrywanie linków w obrazkach
- ukrywanie linków w skryptach

## Spamowanie Linków, c.d.

Przykładowe techniki:

- kopiowanie dużej ilości wartościowych linków wychodzących (cel: np. hub score, etc.)
- tworzenie wielu kopii innych stron z dodanymi (ukrytymi) linkami do strony docelowej (ang. honey pot) (cel: np. authority score, etc.)
- tworzenie dużych ilości linków ze stron z kategorii “dostępne” do stron docelowych (np. spam blogowy - ang. blog spam - **poważny problem**)
- kupowanie przeterminowanych domen i zapełnianie ich stronami docelowymi
- tworzenie **farm linków**
- wymiana linków (ang. link exchange) z niezwiązanymi witrynami

Ostatnie 2 przykłady są szczególnie ciekawe z naukowego punktu widzenia. Pojawia się tam nietrywialna matematyka i elementy teorii gier (w ostatnim).

## Inne Techniki

Popularne techniki spammerskie wykorzystują też właściwości protokołu HTTP:

- Używanie nazw URL podobnych do innych, znanych, ale nieprawidłowo zapisanych
- Maskowanie (ang. cloaking) - pokazywanie innej zawartości crawlerowi (do indeksowania) a innej (docelowa strona) regularnym użytkownikom (przeglądarkom). Niektóre crawlery obchodzą ten problem podając się za zwykłe przeglądarki - ale to z kolei nie jest w pełni "fair" ze strony wyszukiwarek (poza tym crawler podający się za przeglądarkę może mieć czasami mniejszy dostęp do pewnych zasobów)
- Natychmiastowe przekierowywanie (ang. redirection) do innej strony. Strona pierwotna (ang. doorway page) jest i tak indeksowana, ale użytkownik nigdy jej nie zobaczy. Trafi natomiast na stronę docelową. (używany jest np. meta znacznik "refresh" albo skrypty)

## Prosty spam

The screenshot shows a web browser window with the following content:

Best deal for car hire discount, LOW COST CHEAP CAR HIRE. The lowest cost self drive rental in the UK. DI

File Edit View Go Bookmarks Tools Help None My Yahoo! SK posts com Ecosofia com


http://www.carhire.ndo.co.uk/

Tejedores del Web Spam Classification http://local...ollection=1 Best deal for car ...

**[cheap car hire call center \[details here\]](#) or complete our simple [cheap car hire enquiry form \[here\]](#) and we will call you back.**

[Cheap Auto Rental] [Cheap Airport Parking] [Cheap Travel Insurance] [Cheap Foreign Currency]  
[Cheap Flight Tickets] [Cheap Hotel Deals] [Cheap Rentals] [Cheap Packages Holidays] [Cheap Weekend Breaks]

Indexed by [Linksmatch](#)  
[Terms & Conditions](#), [Privacy Policy](#).  
[cheepcar.co.uk](#) copyright [cheeptravel Limited](#)@  
[cheeptravel Limited](#)@ part of the DHD Group Limited

 **RINGTONES, LOGOS & PICTURE MESSAGES ?**  
**U CAN GET THEM @ [RECMONGOOSE.COM](#)**

DISCOUNTED CAR HIRE IN THE UK. For the best deal on CHEAP car hire rental in the United Kingdom, visit our UK DISCOUNT SELF DRIVE feature. Guaranteed discount off normal self drive rates. DISCOUNTED CAR HIRE IN THE UK. For the best deal on CHEAP car hire rental in the United Kingdom, visit our UK DISCOUNT SELF DRIVE feature. Guaranteed discount off normal self drive rates. DISCOUNTED CAR HIRE IN THE UK. For the best deal on CHEAP car hire rental in the United Kingdom, visit our UK DISCOUNT SELF DRIVE feature. Guaranteed discount off normal self drive rates. DISCOUNTED CAR HIRE IN THE UK. For the best deal on CHEAP car hire rental in the United Kingdom, visit our UK DISCOUNT SELF DRIVE feature. Guaranteed discount off normal self drive rates. ....

Scripts Currently Forbidden [<script>: 5] [+F+P: 0]

Done Proxy: None Adblock



## Ukryty tekst

1 Stop Poker Games - Mozilla Firefox

File Edit View Go Bookmarks Tools Help None My Yahoo! SK posts com Ecosofia com Diggs

http://www.1stoppoker.com/games.html

Web Spa... webspam... http...xhtml mp3 ware... 1 Stop P... Hidden Text 'poker po...

**Play Now!**  
**Million Dollar**  
Guaranteed  
Tournament

Sign Up for 1 Stop's Newsletter!

Your E-Mail:  
  
Subscribe!

Great Poker Tips, #8

There are only about 20 hands that are strong enough to play from an early position. Players are making a big mistake if they play weak or marginal hands without giving consideration to their position.

— Bill Burton. *Get the Edge at Low Limit Texas Hold'Em*. Bonus Books, 2002

Want the edge?  
Get the book!

**THE EDGE AT LOW LIMIT**

**Texas hold 'em** (or simply **hold 'em** or **holdem**) is the most popular of the **community card** poker games. It is the most popular **poker variant** played in **casinos** in the western **United States**, and its **no-limit** form is used in the main event of the **World Series of Poker** (abbreviated **WSOP**), widely recognized as the world championship of the game.

**Seven-card stud** is a **poker variant**. Until the recent increase in popularity of **Texas hold 'em**, Seven-card stud was the most popular poker variant in home games across the **United States**, and in **casinos** in the eastern part of the country.

**Omaha hold 'em** (or **Omaha holdem** or simply **Omaha**) is a **community card poker** game based on **Texas hold 'em**. It was originally created as a high-hand only game, but a **high-low split** variant called "Omaha eight-or-better" has also become popular.

**Five-card draw** is often the first **poker variant** learned by most players, and is very common in home games although it is now rare in **casino** and **tournament** play. The **lowball** variations make more interesting games and are more commonly played in casinos. Two to eight players can play.

Source: [Wikipedia](#), the free encyclopedia.


poker poker poker poker poker  
poker poker poker poker poker poker poker poker  
poker poker poker poker poker poker poker  
poker poker poker poker poker poker poker poker  
pokerpoker poker poker poker poker poker poker

Scripts Currently Forbidden [<script>: 1] [+F+P: 0]

Done Adblock Proxy: Non

## Wyszukiwarka?

Bookmark Home  
Page Home



**Top Searches:**

- » Acne
- » Weight Loss Pills
- » Debt Consolidation
- » Loan
- » Domain Names
- » Advertising
- » Online Pharmacy
- » Home Loan
- » Dedicated Server
- » Car Rental
- » Adipex
- » Levitra
- » Online Poker
- » Work At Home
- » Propecia
- » Consolidate Debt
- » Mortgage Rates
- » Online Craps
- » Vegas Casinos
- » Buy Ionamin



java soft	php script	top soft	java script	MP3
-----------	------------	----------	-------------	-----


**Top Web Results**

Results 1-16 containing "sports book"

1. **Place Your Bet with #1 Sports Betting Site Online**  
Kentucky Derby, NBA, MLB, NHL and all other sports betting and odds. Place a full range sportsbook in North America  
<http://www.sportsinteraction.com>
2. **AnteUp GamblingLinks.com - Safe Online Casinos**  
Links to safe and secure online casino gambling and sports betting including reviews, ne  
<http://gamblinglinks.com>
3. **Free Casino Bonuses. Links To the Best Casinos**  
Get \$20 - \$500 in Free Chips. Most popular casino games with great graphics. Play for fun rules and strategy. Links to the Best Casinos  
<http://www.fastfreecash.net>
4. **AnteUp GamblingLinks.com - Safe Online Casinos**

# Fałszywa wyszukiwarka

→ Bookmark → Home Page → Home



**SOFT SEARCH**

**Top Searches:**

- » Canadian Pharmacy
- » Debt Consolidation
- » Online Loan
- » Diet
- » Credit Reports
- » Online Poker
- » Xenical
- » Buy Ionamin
- » Diet Pills
- » Online Craps
- » DirecTV
- » Life Insurance
- » Dedicated Server
- » Car Insurance
- » Buy Phentermine
- » Debt
- » Weight Loss Pills
- » Pay Day Loans
- » Home Loan
- » Refinance



lava soft

php script

top soft

java script

MP3

## Top Web Results

Results 1-16 containing "1293kasd132ka0sd1kj239asd123"

1. **A Real Work At Home Business Opportunity!**  
Free Home Business Match Up Service! We have helped 1000's of people make \$5,000 per month.  
<http://gozing.directtrack.com/z/1198/CD2127/>
2. **Exotic Holiday - Find Your Love**  
Exotic holiday is great way how to find love when you travel. Meet new people. Meet new friends.  
<http://www.exotic-holiday.co.uk/>
3. **Image, Photo, Digital, Video and Movie software**  
Find quality image management & digital asset software for your business. Also search engines optimization software.  
<http://www.enterprise-software.co.uk>
4. **Renting a Birthday Party Limousine is Sexy**  
What better way to surprise your loved one on their special day than with a birthday party?  
<http://partybusrental.info>

# Jedyną treścią reklamy

The screenshot shows a Mozilla Firefox browser window with the following details:

- Address Bar:** <http://www.home-security-webpage.com/home-security-system-separate-blasts-kill-ne>
- Page Title:** Home Security Webpage
- Search Results:**
  - Ads by Google:**
    - [Alarm Systems](#): Looking to find alarm systems? Visit our alarm systems guide. OnlyAlarmSystems.com
    - [Security Systems](#): Selected Security System Deals Find Exactly What You Want Today www.Security-Systems.in
    - [Centurion Wireless System](#): Panic Alarm System for Public Facilities and Courthouses. www.stoptechhd.com
  - Archived Entry:**
    - Post Date:** Tuesday, Nov 22nd, 2005 at 2:03 pm
    - Category:** Uncategorized
    - Do More:** You can trackback from your own site.
  - Ads by Google (continued):**
    - [Prevent Home Burglary](#): Home burglary is rampant. Read all about security systems. www.for-the-touchdown
    - [Security Industry News](#): Latest on CCTV, loss prevention, access control & more for pros
- Main Content:**
  - Uncategorized** 22 Nov 2005 02:03 pm
  - Home security system - Separate Blasts Kill Nearly 100 in Iraq**
  - Separate Blasts Kill Nearly 100 in Iraq**
  - Washington Post - By Ellen Knickmeyer and Naseer Nouri Washington Post Foreign Service Saturday, November 19, 2005; Page A01 BAGHDAD, Nov. AP) Video Security Video Shows Huge Explosion Video from a security camera at the Hamra Hotel in Baghdad look at the fallen troops/home towns, ages, service categories and other
  - Rood girl's game of strip
- Search Bar:** Find: filters Find Next Find Previous Highlight all Match case
- Status Bar:** Done Disabled Proxy: None

## Website design, management, marketing and promotion

If you are searching for any of the following topics:

- ◆ [Website design, management, marketing and promotion.](#)
- ◆ [Website design, management, marketing and promotion resources.](#)
- ◆ [Website design, management, marketing and promotion related topics.](#)
- ◆ [Website design, management, marketing and promotion services.](#)

Look No further. You'll find it at [Website design, management, marketing and promotion](#)!

Website design, management, marketing and promotion is the key to your needs. You're one step ahead with Dry Media.

Website design, management, marketing and promotion brought to you by Dry Media, the leaders in this field.

At the [Website design, management, marketing and promotion web site](#), you'll discover an easy to use, information packed source of data on Website design, management, marketing and promotion.

[Click Here to Learn More about Website design, management, marketing and promotion.](#)

# Przykłady przekierowań w javascript

## Proste przekierowanie

```
<script>
document.location="http://www.topsearch10.com/";
</script>
```

## Ukryte przekierowanie

```
<script>
var1=24; var2=var1;
if(var1==var2) {
    document.location="http://www.topsearch10.com/";
}
</script>
```

(Przykłady wzięte z Chellapilla et al. "A taxonomy of JavaScript redirection spam", 2007)

## Mocno ukryte przekierowanie

```
<script>
var a1='win',a2='dow'",a3='loca',a4='tion.',
a5='replace',a6='('http://www.top10search.com/')';
var i,str='';
for(i=1;i<=6;i++)
{
  str += eval('a'+i);
}
eval(str);
</script>
```

(Przykłady wzięte z Chellapilla et al. "A taxonomy of JavaScript redirection spam", 2007)

# Bardzo mocno ukryte przekierowanie

## Zakodowany javascript

```
<script>
var s = "%5CBEOD%5C%05GDHJ_BDE%16...%04%0E";
var e = ", i;
eval(unescape('s%eDunescape%28s%29%3Bfor...%3B'));
</script>
```

(Przykłady wzięte z Chellapilla et al. "A taxonomy of JavaScript redirection spam", 2007)



## Wytyczne dla twórców stron

Każda większa wyszukiwarka publikuje swoje wytyczne, których autorzy stron powinni przestrzegać aby nie zostać uznanymi za spamerów.

Do najważniejszych, wspólnych reguł należą:

- zakaz automatycznego nieautoryzowanego odpytywania (lub klikania) (badanie algorytmów rankingowych, atakowanie kampanii reklamowych konkurentów, etc.)
- unikanie powtarzania/kopiowania treści na wielu stronach, domenach
- unikanie przekierowywania i maskowania
- niestosowanie ukrytych linków i tekstu
- nieuczestniczenie w programach wymiany linków
- dostarczanie wysokiej jakości, oryginalnej zawartości związanej ściśle z tematyką strony/witryny
- unikanie linków do i z (!) “podejrzanych” dokumentów

Wobec podmiotu łamiącego **wyciągane są konsekwencje** w przypadku wykrycia przez wyszukiwarkę złamania reguł.

# Techniki wykrywania spamu - podstawowe techniki tekstowe

Automatyczne wykrywanie najprostszych technik powtarzania i ukrywania tekstu nie nastręcza dzisiaj większych trudności. Stosuje się techniki statystyczne i probabilistyczne.

Nieco więcej problemów jest z technikami “zaśmiecania” czy “rozpuszczania”. Powstają coraz doskonalsze modele języka. Wykrywanie “zaśmiecania” można zaimplementować np. jako badanie tematyki dokumentu. Rozpuszczanie można wykryć np. za pomocą HMM (Ukrytych Modeli Markowa) wyższych rzędów, lub badania relacji sąsiadujących wyrazów.

## Wykrywanie spamu - techniki oparte na linkach

Analiza statystyczna rozkładów stopni wejściowych i wyjściowych.

Analiza rozkładu wartości PageRank.

Specjalne algorytmy: TrustRank i Anti-TrustRank, BadRank, etc.

Wykrywanie “farm” linków - bardziej skomplikowane metody kombinatoryczne i statystyczne.

Wykrywanie bardziej zaawansowanych struktur spamowych i “wymian linków” (ang. “spam alliances”) - elementy teorii gier

Wykrywanie “spamów odnośników” (ang. anchor text spam) - bardzo poważny problem w związku z blogami - badanie stopnia niezgodności modelu języka (w otoczeniu źródła linku i w dokumencie docelowym)

## Techniki Wykrywania Spamu c.d.

Wykrywanie spamu na etapie ściągania dokumentów:

Prowadzenie i stałe uaktualnianie “czarnych list”

Analiza rozkładów wielkości hostów, domen, poddomen, etc.

Analiza nazw URLi, hostów, plików i struktury witryn

Wykrywanie przekierowań

## Stan Obecny

W 2007 został przygotowany ogólno-dostępny zbiór dokumentów z angielskiego WWW<sup>1</sup>, wraz z etykietami, do testowania najnowszych algorytmów wykrywania spamu. Dostępne jest też prawie 300 atrybutów (!) policzonych dla każdego z ok. 11 000 hostów.

Bardzo dobre efekty w wykrywaniu spamu daje ostatnio stosowanie technik uczenia na grafach (ang. stacked graphical learning) biorące pod uwagę sąsiedztwo dokumentów.

Najnowsze i najlepsze obecnie (2007) techniki wykrywania spamu stosują analizę “ekonomiczną” zawartości dokumentów WWW, stosując narzędzia związane z przygotowywaniem kampanii reklamowych (Google AdWords, Yahoo! Mindset, Microsoft AdCenter OCI, etc.) i dają obiecujące rezultaty.

---

<sup>1</sup>zbiór wraz z atrybutami jest dostępny m.in. w katedrze SI PJWSTK

## Wykrywanie Spamu - podsumowanie

Należy pamiętać, że rozwojowi technik wykrywania spamu towarzyszy ciągły postęp w technikach spammerskich (“wyścig zbrojeń”).

Najbardziej wartościowe są te techniki wykrywania, które są trudne do “oszukania” nawet, gdy spamerzy poznają ich istotę (np. oparte na zasadach ekonomii, albo wymagające posiadania unikatowych danych)

Należy też podkreślić, że istotną barierą w wykrywaniu spamu jest bariera złożoności czasowej algorytmów <sup>2</sup>

Ogromnie ciekawe zastosowania dla Web Mining i Sztucznej Inteligencji - wiele otwartych problemów.

---

<sup>2</sup>praktycznie, rzędu  $o(n^2)$  (“o małe”), gdzie  $n$  to liczba wyrazów w dokumencie, wierzchołków w grafie, etc.

## Podsumowanie - wyzwania

Rynek wyszukiwarek oferuje fascynujące problemy naukowe i techniczne do rozwiązania

- Młody rynek o wartości **kilku miliardów dolarów** (obecnie), gwałtownie rosnący
- Fascynujący styk Informatyki, Teorii Informacji, Sztucznej Inteligencji, Matematyki, Mikroekonomii (z teorią gier i odwróconą teorią gier) a nawet Socjologii czy Psychologii
- Wiele otwartych problemów o praktycznym wymiarze (i realnej wartości) np. modele wyceny reklam
- Wciąż młoda dziedzina - można jeszcze w niej sporo osiągnąć!

## Na zaliczenie tego wykładu:

- 1 Spam wyszukiwarkowy - definicja i motywacja
- 2 Spam a pozycjonowanie
- 3 Rodzaje technik spammerskich
- 4 Metody wykrywania spamu



Dziękuję za uwagę