Wyszukiwanie i Przetwarzanie Informacji – Web Spam

Information Retrieval & Search

Irmina Masłowska

irmina.maslowska@cs.put.poznan.pl

http://www.cs.put.poznan.pl/imaslowska/wipi/

Web spamming

- **web spamming** – umyślne działanie człowieka mające na celu oszukanie algorytmów rangujących wyszukiwarek internetowych, tak by oceniały niektóre strony wyżej niż na to zasługują (wyższa pozycja na SERPs – *search engine results pages*)

Web spamming

- **web spamming** – umyślne działanie człowieka mające na celu oszukanie algorytmów rangujących wyszukiwarek internetowych, tak by oceniały niektóre strony wyżej niż na to zasługują (wyższa pozycja na SERPs)
- 8%-13% indeksowanych stron (2006)
- szacowane straty finansowe spowodowane spamem: 2005 – 50 miliardów $, 2009 -130 miliardów $

N. Spirim, J. Han: Survey on Web Spam Detection: Principles and Algorithms

Web spamming

- **web spamming** – umyślne działanie człowieka mające na celu oszukanie algorytmów rangujących wyszukiwarek internetowych, tak by oceniały niektóre strony wyżej niż na to zasługują (wyższa pozycja na SERPs)

web spamming  search engine optimization (SEO)

(*pol*. pozycjonowanie)

Typy spamu

- Content spam
- Link spam
- Click spam
- Cloaking & Redirection

Content spamming

Content spamming (lub *term spamming*) to umieszczanie nierzetelnej informacji w:

- tytule strony
- meta-tagach
- sekcji “body”
- tekście hiperłącz
- adresach URL
- ...

Keyword stuffing

aby dopasować treść stron do zapytań użytkowników wyszukiwarek stosuje się „upychanie” odpowiednich słów kluczowych we wszystkich możliwych tekstowych elementach stron

np. opisach alternatywnych:

```
<img src="obrazek.jpg" alt="tu powinien być opis
tego, co faktycznie znajduje się na obrazku"/>
```

Keyword stuffing

- Wielokrotne powtarzanie wybranych słów kluczowych, aby zwiększyć dopasowanie strony do wybranych zapytań użytkowników
- Wrzucanie wielu słów kluczowych luźno związanych z treścią strony, aby zwiększyć dopasowanie tej strony do większej liczby zapytań użytkowników

Keyword stuffing

W przypadkach skrajnego spamu cała treść strony może być generowana sztucznie:

- poprzez skopiowanie treści innych stron/-y o podobnej tematyce
- poprzez wstawianie do spam-strony zdań skopiowanych z wielu innych różnych stron

W obu przypadkach stosuje się tzw. *web scraping* (*web harvesting*) aby pozyskać naturalnie wyglądającą treść bazową

Spamowanie odnośników

- dotyczy manipulowania hiperłączami, zarówno wychodzącymi z danej spam-strony (*outbound links*), jak i wskazującymi na daną spam-stronę (*inbound links*, *backlinks*)

Linki wychodzące

Reboot created 10 new websites each targeting the same keyword, only half of which included links to high authority sites. After five months it was concluded that, "*Outgoing relevant links to authoritative sites are considered in the algorithms and do have a positive impact on rankings*"

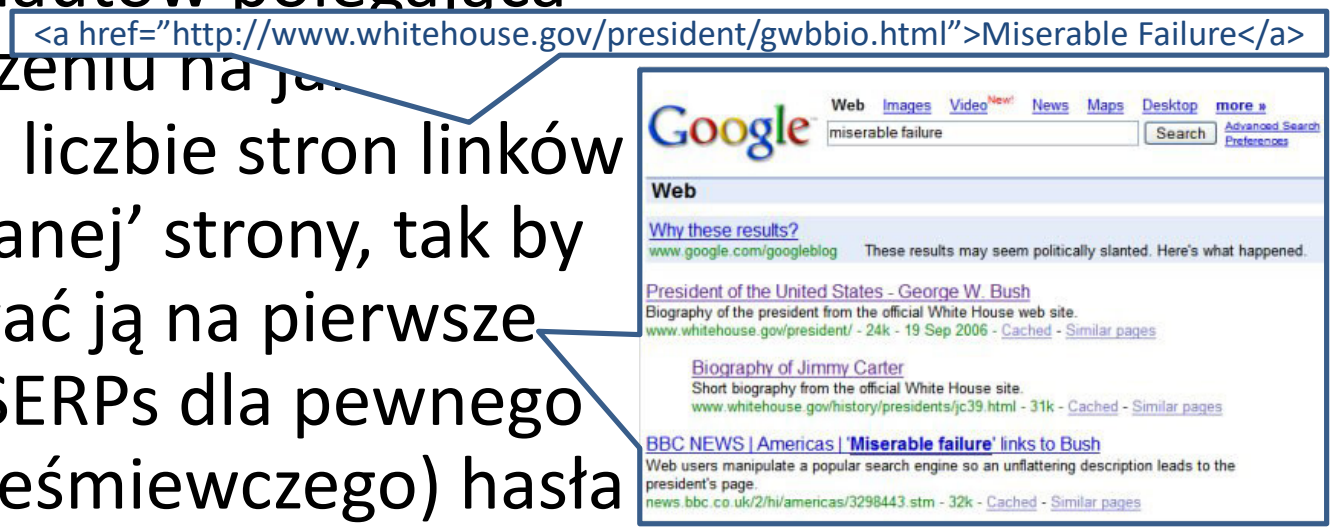
https://searchenginewatch.com/2016/11/02/

Spamowanie odnośników wychodzących

- linki wychodzące – dość łatwo umieścić na własnej spam-stronie dużą liczbę linków do ważnych stron (np. *directory cloning*), aby podnieść jej *hub score*
- *Google bomb* – zorganizowana akcja internautów polegająca na umieszczeniu na jak największej liczbie stron linków do 'atakowanej' strony, tak by wypromować ją na pierwsze miejsce w SERPs dla pewnego (zwykle prześmiewczego) hasła

Spamowanie odnośników wychodzących

- linki wychodzące – dość łatwo umieścić na własnej spam-stronie dużą liczbę linków do ważnych stron (np. *directory cloning*), aby podnieść jej *hub score*
- *Google bomb* – zorganizowana akcja internautów polegająca na umieszczeniu na ja... największej liczbie stron linków do 'atakowanej' strony, tak by wypromować ją na pierwsze miejsce w SERPs dla pewnego (zwykle prześmiewczego) hasła



Spamowanie odnośników wskazujących

linki wchodzące – manipulowanie linkami wskazującymi z innych (obcych) stron jest trudniejsze, ale istnieją mniej lub bardziej kosztowne metody:

- wstawianie linków na daną spam-stronę do komentarzy zamieszczanych w serwisach społecznościowych, blogach, dyskusjach na forach, recenzjach (czyli w tzw. *user-generated content*) lub wiki (*wiki spam*)
- dodanie linków na daną spam-stronę do katalogów stron (*web directories spamming*)
- kupowanie linków (*paid links*)

Spamowanie odnośników wskazujących

- organizacja grupy wzajemnej wymiany linków – kooperacja między spamerami (*link exchange*)
- utworzenie własnej spam-farmy, co pozwala na utworzenie dowolnej struktury linków, lecz wymaga kontroli nad większą liczbą witryn
- tworzenie stron zwanych *honey pot*, ukrywających linki do spam-strony, której ranking chcemy polepszyć. Atrakcyjność stron *honey pot* wynika z zamieszczenia na nich informacji użytecznych dla wielu użytkowników (np. list FAQ lub dokumentacji dla popularnych narzędzi)

Click spamming

Spamowanie kliknięć

- *Clik spamming* dotyczy technik generowania zapytań do popularnych wyszukiwarek internetowych, aby następnie wybierać (click) daną spam-stronę z listy wyników SERP, a przez to ją wypromować symulując zainteresowanie prawdziwych użytkowników
- Click spam ma na celu podniesienia rankingu strony
- podobnie działa tzw. *click fraud* – zjawisko nieuczciwych, czy fałszywych kliknięć w link sponsorowany (lub inną formę reklamy) rozliczany w modelu PPC. Ma ono na celu wygenerowanie wyższych kosztów konkurencyjnym reklamodawcom i „zużycie" opłaconych przez nich wyświetleń reklam

Ukrywanie zawartości

Użycie (prawie) tego samego koloru tekstu co tła

```
<body background = white>
      <font color = close-to-white>spam items</font>
      ...
</body>
```

Zamieszczenie ukrytych hiperłącz

```
<a href="spam_target.html"><img src="blank.gif"></a>
```

Ukrywanie zawartości

Użycie (prawie) tego samego koloru tekstu co tła

```
<body background = white>
      <font color = close-to-white>spam items</font>
      …
</body>
```

Zamieszczenie ukrytych hiperłącz

```
<a href="spam_target.html"><img src="blank.gif"></a>
```

Stosowanie skryptów do ukrycia części stron

Spam hiding techniques

Cloaking (maskowanie)

Prezentowanie wyszukiwarce internetowej treści odmiennej od zawartości prezentowanej czytelnikom

Identyfikacja robotów indeksujących na podstawie predefiniowanych list adresów IP wyszukiwarek lub też analizy nagłówka *user-agent*

Redirection (przekierowywanie)

Strona prezentowana robotom nie jest widoczna dla użytkowników, np. dzięki użyciu skryptów lub meta-tagu 'refresh' (`Refresh: 0; url=address`)

Combating spam

Wyszukiwarka Google w swoich oryginalnych założeniach była dość odporna na ówczesne metody spamowania:

- PageRank dobrze radził sobie z technikami spamowania zawartości, choć był w pewnym stopniu podatny na techniki manipulowania linkami wchodzącymi
- wykorzystanie w charakterze termów indeksujących słów z tekstów łącz (ang. *anchor text*) linków zewnętrznych do określania tematycznej zawartości wskazywanych stron pozwalało ocenić tematykę strony w oderwaniu od jej własnej treści (ale *Google bombs*)

Combating spam

‘Jawny’ spam (keyword stuffing, ukrywanie elementów, przekierowania, linki nieorganiczne, niedopasowane do tematyki strony, itp.) dość łatwo identyfikować

Wykorzystanie skryptów komplikuje sprawę, gdyż ich wykonywanie/analizowanie przez serwisy wyszukujące może być zbyt czaso-/kosztochłonne

Combating spam – dynamic content

Zalecenia Google'a (m.in. odnośnie użycia JavaScript)

- "Don't cloak to Googlebot"
- "Google supports JavaScript to some extent (titles, description & robots meta tags, structured data, and other meta-data)"
- "Use the rel=canonical attribute" (*duplicate content*)

Pełna lista zaleceń skompilowana na podstawie wypowiedzi Johna Muellera w marcu 2016 dostępna na

https://blog.seoprofiler.com/googles-john-mueller-google-indexes-javascript-sites/

Can Google Properly Crawl and Index JavaScript Frameworks? A JavaScript SEO Experiment

https://www.elephate.com/blog/javascript-seo-experiment/
https://www.elephate.com/blog/everything-you-know-about-javascript-indexing-is-wrong/

Algorytmy segmentacji dokumentów

Warto różnicować ważności słów kluczowych i linków występujących na stronie – zależności od ich położenia

Automatyczna identyfikacja obszarów dokumentu tekstowego pomoże podzielić go na partie ważniejsze – na których skupia się użytkownik i partie poboczne – gdzie nie powinny się znajdować żadne kluczowe informacje

Najczęściej podejrzane treści czy linki będą umieszczane w takich miejscach dokumentu, które nie rzucają się w oczy użytkownikom (np. na samym dole stron typu *honey pot* lub stron uczestniczących we wzajemnej wymianie linków), aby uniknąć zgłoszenia spamu do wyszukiwarek (*spam report*)

Combating spam – links

Algorytmy grafowe (web structure mining)

Wykorzystanie obserwacji, że wartościowe strony i spam-strony tworzą pewne odrębne obszary sieci Web

Strony wysokiej jakości rzadko wskazują na spam-strony, jednak spam-strony mogą wskazywać na strony wartościowe

Linki wchodzące ze stron o słabej reputacji mogą oznaczać kłopoty (TrustRank, Google Penguin, ‘manual’ actions)

Google's manual actions

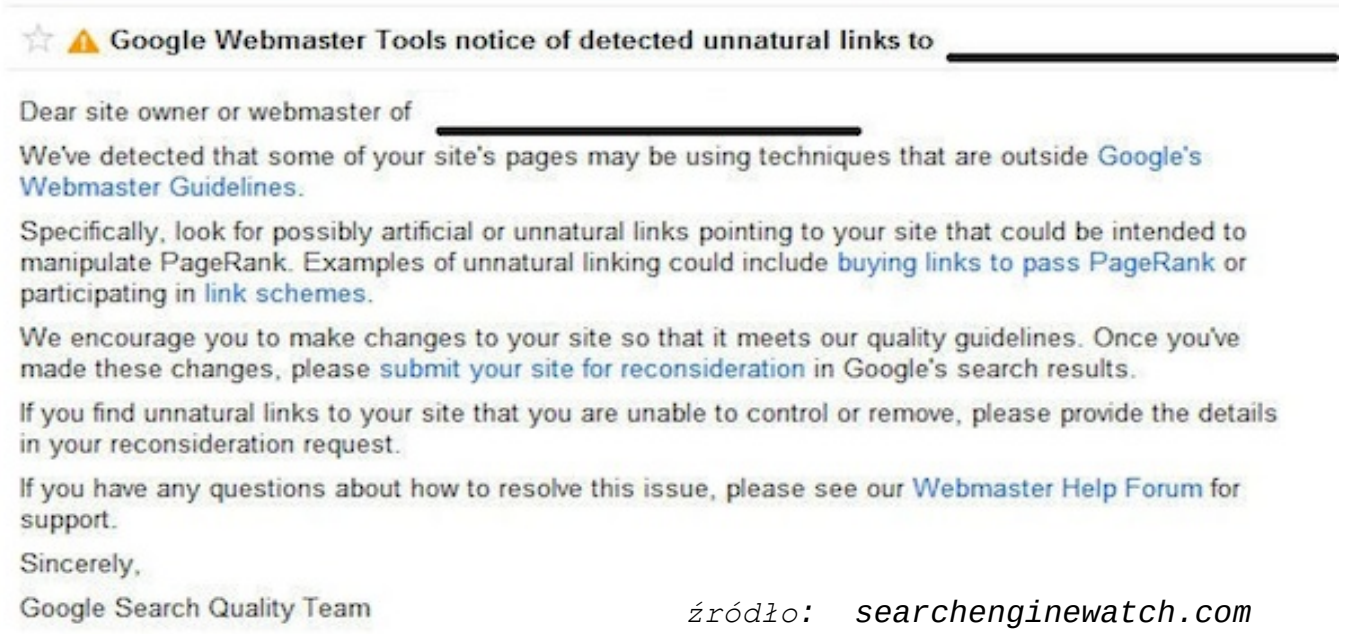
W kwietniu 2012 Google zapowiedział obniżenie obniżenia wartości tzw. 'nieorganicznych' (nienaturalnych) linków:

- pochodzących z „farm linków"
- pochodzących z sieci wymiany linków
- wszelkich innych, które noszą znamiona linków płatnych (w szczególności otagowanych tekstem w postaci popularnych słów kluczowych)

Google's manual actions

W kwietniu 2012 Google zapowiedział obniżenie obniżenia wartości tzw. 'nieorganicznych' (nienaturalnych) linków:

- pochodząc
- pochodząc
- wszelkich i (w szczegó słów klucz



☆ ⚠ **Google Webmaster Tools notice of detected unnatural links to** [redacted]

Dear site owner or webmaster of [redacted]

We've detected that some of your site's pages may be using techniques that are outside Google's Webmaster Guidelines.

Specifically, look for possibly artificial or unnatural links pointing to your site that could be intended to manipulate PageRank. Examples of unnatural linking could include buying links to pass PageRank or participating in link schemes.

We encourage you to make changes to your site so that it meets our quality guidelines. Once you've made these changes, please submit your site for reconsideration in Google's search results.

If you find unnatural links to your site that you are unable to control or remove, please provide the details in your reconsideration request.

If you have any questions about how to resolve this issue, please see our Webmaster Help Forum for support.

Sincerely,

Google Search Quality Team

źródło: searchenginewatch.com

Google's manual actions

W październiku 2012 Matt Cutts poinformował o nowym narzędziu dla webmasterów (*tool to disavow links*), które umożliwia przekazanie do Google w prostym pliku tekstowym listy adresów serwisów (lub pojedynczych stron), z których linki do naszych stron uznajemy za niewartościowe

Przykładowa treść pliku:

```
# Contacted owner of spamdomain1.com on 7/1/2012 to
# ask for link removal but got no response
domain:spamdomain1.com
# Owner of spamdomain2.com removed most links, but missed these
http://www.spamdomain2.com/contentA.html
http://www.spamdomain2.com/contentB.html
http://www.spamdomain2.com/contentC.html
```

Combating spam – backlinks

Google's manual actions

W październiku 2012 Matt Cutts poinformował o nowym narzędziu dla webmasterów (*tool to disavow links*), które umożliwia przekazanie do Google w prostym pliku tekstowym listy adresów serwisów (lub pojedynczych stron), z których linki do naszych stron uznajemy za niewartościowe

Z punktu widzenia Google'a bardzo prosta i tania metoda zaangażowania rzeszy darmowych redaktorów (*human editors*) do wskazania spamerskich serwisów/obszarów sieci Web

Combating spam as a ML task

Zasadniczo każda technika spamowania skupia się na jednym lub wielu aspektach wykorzystywanych przez algorytmy rangujące wyszukiwarek, co może skutkować nienaturalną charakterystyką takich stron

W ogólności heurystyki detekcji spamu poszukują statystycznych anomalii w cechach stron/witryn, głównie obserwowalnych z poziomu wyszukiwarki

Detekcja spamu sprowadza się obecnie do problemu **klasyfikacji** $\Leftarrow$ wysoka skuteczność metod uczenia maszynowego ML

Combating spam as a ML task

Zasadniczo każda technika spamowania skupia się na jednym
algorytm
nienatu

W ogóln
statysty
głównie

Detekcja spamu wadza się obecnie do problemu **klasyfikacji** ⇐ wysoka skuteczność metod uczenia maszynowego ML

"Web spam classification: a few features worth more", M.Erdélyi, A. Garzó, and A. A. Benczúr , ACM Press 2011

>>*we investigate how much various classes of* ***Web spam features, some requiring very high computational effort****, add to the classification accuracy. We realize that* ***advances in machine learning****, an area that has received less attention in the adversarial IR community,* ***yields more improvement than new features*** ***and result in low cost yet accurate spam filters***<<

Użyte techniki ML*: LogitBoost*, *random forests*

Obecnie popularne: techniki *deep learning*

Combating spam as a ML task

Cechy stron, które mogą być przydatne w klasyfikacji:

- liczba unikatowych słów na stronie (spam-strony zwykle zawierają ich więcej),
- średnia długość słowa (dla j. angielskiego =5), która jest często odmienna dla nienaturalnych tekstów,
- liczba słów w tytule strony (zwykle wyższa dla spam-stron),
- odsetek stopwords (często niższy dla nienaturalnych tekstów),
- procent zawartości widocznej (często niższy dla spam-stron),
- ilość tekstu w opisach łącz (większa dla spam-stron),
- TLD, IP,
- podatność na kompresję,
- ...

Wykorzystanie informacji i powiązań z serwisami społecznościowymi:

Table 1: Feature importance analysis.

Top 10 Features	Rank
fraction of tweets containing URLs	1
average number of URLs per tweet	2
fraction of followers per followees	3
average number of users mentioned per tweet	4
number of followees	5
number of tweets with spam words	6
fraction of tweets that are reply messages	7
number of followers	8
average number of #hashtags per tweet	9
median number of users mentioned per tweet	10

źródło: Mutually Reinforcing Spam Detection on Twitter and Web Nikita Spirin, 2011

Assesing the quality of sites

Rozwój metod ML przyczynił się do lepszej identyfikacji spamu przez wyszukiwarki – takie strony są wyłączane z rankingu

Możliwość zastosowanie metod ML na dużą skalę wykorzystano do okresowej oceny regularnych stron pod kątem ich szeroko pojętej jakości (Google's *Panda* aka *Farmer* – wprowadzony w lutym 2011, a od stycznia 2016 działający jako integralna część mechanizmu rangującego Google'a)

Efekt: wyraźnie zauważalne zmiany pozycji na listach SERPs wg Google pierwsza edycja Pandy wpłynęła na wyniki ok. 12% zapytań kierowanych do ich wyszukiwarki

Kolejne zmiany algorytmu rangującego Google'a: *Hummingbird* (2013) i *RankBrain* (wprowadzany od 2015)

Assesing the quality of sites

8 major Google ranking signals in 2017

Cechy stron, wykorzystywane w ocenie jakości:

- Backlinks (*link score*, *anchor text relevance* - but not too much of it – exact match keywords in links may get punished)
- Content (*keyword usage*, *length*, *comprehensiveness*)
- Technical SEO (*page speed*, *mobile-friendliness*)
- User experience (SERP CTR – *clickthrough rate*)

Źródło: https://searchengineland.com/8-major-google-ranking-signals-2017-278450