Spam-szűrés

Zongor Attila

Bevezetés

- Mi a spam?
- Miért kell szűrni a spameket?
- Hogyan védekezhetünk a spamek ellen?
- Hogyan működnek a spam szűrök?

Mi s spam?

- A spam legáltalánosabban kéretlen és tömeges e-mail üzenetet jelent.
- A címzett személye érdektelen, mivel az üzenetet változatlan formában lehetne sok más felhasználónak is küldeni.
- A címzett nem engedélyezte a levél küldését.
- A levél küldése és fogadása lényegesen nagyobb előnyt jelent a küldő számára

Miért kell szűrni a spamet?

- Nem szükség, viszont rengeteg időt takaríthatunk meg.
- Nem kell minden egyes levelet, mivel manapság naponta több száz spam is érkezhet postaládánkba.

Hogyan védekezhetünk a spamek ellen?

- Ne adjuk meg e-mail címünket ismeretlen helyen.
- Ne használjuk kedvenc email címünket elektronikus listákon.
- Nem nézem meg “EGYSZER AZ ÉLETBEN VAN ILYEN AKCIO” leveleket, ugyanis a levél megnyitásával a küldő megkapja az e-mail címünket amit továbbad…
- Használjuk a szolgáltató spam-szűrő rendszerét.

Hogyan működnek a spam szűrők?

- Statisztikai alapon:
  - Elemzik a levélek szövegét
  - Elemzik a levél fejlécét
  - Pár dolog ami spamre utal:
    - Nincs neve a feladónak, csak email címe.
    - Nem vagyunk közvetlen címzettek, azaz más is megkapta a levelet.
    - Nem normális SMTP szerverrel küldték a levelet, hanem helyi gépről.

Hogyan működnek a spam szűrők?

- Egy tipikus szűrő az alábbi séma szerint épül fel:

Ábra 1. Spam szűrő felépítése

Hogyan működnek a spam szűrők?

- Az ellenőrzés első szintje: a white/black list
  - A feladó címe vagy domain címe alapján történik a kiértékelés.
    - Ha white szerepel a cím a levél minden további ellenőrzés nélkül továbbításra kerül.
    - Ha black listán szerepel, akkor elutasításra kerül.
- A következő lépésben a spam szűrő megvizsgálja, hogy a küldő IP címe szerepel-e valamelyik folyamatosan frissítet tiltólidtan(RBL).
  - Ezek a tiltólisták az ismert spamet küldő IP címeket és szolgáltatókat tartalmazzák.

Hogyan működnek a spam szűrők?

- A harmadik lépésben már a levél tartalmát elemzi a szűrő.
  - Leggyakrabban használt módszer: BAYES-szűrésen alapuló döntéshozatal (annak a meghatározása, hogy egy adott levél milyen valószínűséggel spam). A feltételes valószínűségre vonatkozó Bayes tétel alapján vissza lehet vezetni annak a vizsgálatára, hogy a levél szavai milyen gyakorisággal fordulnak elő spam illetve nem spam levelekben.

Bayes szűrők

- Alapprobléma meghatározása, hogy az adott levél mekkora valószínűséggel spam.
- Bayes és a teljes valószínűség tétel alapján meghatározhatjuk.
- Sok levél feldolgozása után meghatározható hogy a szavak mekkora relatív gyakorisággal fordulnak elő spam illetve nem spam levelekben.
- Nagy mintalevél gyűjteményre van szükség…

Bayes szűrők

- Annak a valószínűsége, hogy a viagra szót tartalmazó levél spam (spamicity)
- Ahol r(viagra|spam) a viagra szónak spam levelekben való előfordulási aránya, (viagra|nospam) ugyan ez tiszta levelekben.
- R(spam) a spam minták aránya a tanító mintákban, értelemszerűen r(nospam) =1-P(spam|viagra)

$$P(spam \mid viagra) \approx \frac{r(viagra \mid spam) \cdot r(spam)}{r(viagra \mid spam) \cdot r(spam) + r(viagra \mid nospam) \cdot r(nospam)}$$

Bayes szűrők

- Egy levél minden szavának meghatározza a spamicity-t.
- Kiválasztja a döntő szavakat és ezekre elvégzi az alábbi összesítést:

$$Spamicity = \frac{\prod_i Spamicity_i}{\prod_i Spamicity_i + \prod_i (1 - Spamicity_i)}$$

•Ha 1-hez közelít spam ha 0-hoz nem spam

Felismerés hatásfokának javítása

- Nem szavakat hanem szó-párokat elemzünk.
  - Probléma: nagy adatbázis!
  - Tanító adatbázis nagysága!

ROBS és ROBEX módszer

$$P(spam \mid token) = \frac{robs * robx + P(token \mid spam)}{robs + P(token \mid spam) + P(token \mid notspam) * scalefactor}$$

- Scalefactor – spam és nem spam levelek hányadosa.
- Robs – mennyire számítson a robx és a token valószínűség.
- Robx – mennyi annak a val., hogy egy olyan token ami nem szerepel az adatbázisunkban, spam levélben szerepel.

ROBS és ROBEX módszer

$$P(sum) = \sum_{tokens} \left(\frac{P(token \mid spam)}{P(token \mid spam) + P(token \mid notspam) * scalefactor} \right)$$

- Megadja a spam valószínűségek összegét.

$$robx = \frac{P(sum)}{Count(spam) + Count(notspam)}$$

- Count(spam) –olyan tokenek száma amelyek csak spam levelekben szerepelnek.
- Count(nospam) –olyan tokenek amelyek csak tiszta levelekben fordulnak elő.

Vakriasztások kiküszöbölésének lehetőségei

- Hibák:
  - Spam beengedések.
  - Nem spam kizárása-
- Megtalálni a megfelelő arányt.

Vakriasztások kiküszöbölésének lehetőségei

- A spamnek minősítés határértékének növelése.
- Levél tanítás.
- Nem látott elemek valószínűségének beállítása.
- Minimum és maximum valószínűség.
  - Megadjuk token minimum és maximum valószínűségét.

Felismerési és vakriasztási arány

	DNSBL	Mintakeresés	Heurisztika	Statisztika
Felismerés	0-60%	80%	95%	99%+
Vakriasztás	10%	2%	0.5%	0.1%

- DNSBL – nyilvántartott domain nevekre vonatkozó tiltólista.
- A statisztika alapon működő spam szűrök különböző felhasználók esetén felismerési arány csak:80-90%

Új spam esetén

- DNSBL – mivel a levél tartalma nem befolyásolja…..
- Mintakeresés – program illetve az adatbázis frissítésével tud lépéstartami.
- Heurisztika – hasonló, mint az előző.
- Statisztika – automatikusan képes megtanulni a spam jellemzőit abban az esetben, ha a felhasználó megadja, hogy melyik levél spam és melyik nem spam

Jellemzők szűrése

- Általános jellemzők: kis nagybetűs arany, betű és más karakter aránya.
- Spamekben használt trükkök felfedése: színek használata, szöveg elrejtése, véletlenszerű szövegek elhelyezése, stb..
- Levél feladójának ellenőrzése.
- Levél címzettjének ellenőrzése.
- Levél tárgyának vizsgálata.
- Stb.

Spamekben használt trükkök

- Képekben tárolt szöveg.
- Láthatatlan tinta
- Szavak tördelése több html tag-gel.
- Szöveg elrejtése HTML elé.
- Táblázatok.
- Szavak betűkre tördelése.

Program

- Megvalósítás
- Mit kódoltál,
- Milyen adatokra,
- Milyen eredménnyel