Stimuly pre výskum a vývoj

SPAMIA

Výskum efektivity algoritmov pre inteligentné rozpoznávanie nevyžiadanej elektronickej komunikácie, návrh teoretických modelov nových algoritmov a posúdenie ich účinnosti

Druh projektu: základný
Číslo projektu: MŠ SR – 3709/2010-11

Údaje o projekte

Názov: Výskum efektivity algoritmov pre inteligentné rozpoznávanie nevyžiadanej elektronickej komunikácie, návrh teoretických modelov nových algoritmov a posúdenie ich účinnosti

Riešiteľ: Slovanet, a.s.
Záhradnícka 151
821 08 Bratislava

Zodpovedný riešiteľ: Ing. Erik Lehotský
erik.lehotsky@slovanet.net

Doba riešenia: 01.08.2010 – 31.07. 2013 (36 mesiacov)

Vytvorenie/udržanie pracovných miest vo výskume a vývoji:
2 miesta (3 pracovníci – jeden plný úväzok bol rozdelený na 2 čiastočné)

Hlavný cieľ projektu: Systematický komparatívny výskum vybraných riešení na detekciu nevyžiadanej elektronickej komunikácie a teoretický návrh nových inteligentných algoritmov eliminujúcich nedostatky skúmaných riešení a použiteľných v podmienkach vysokorýchlostných sietí.
Získanie nových poznatkov o účinnosti algoritmov pri rozpoznávaní nevyžiadanej elektronickej komunikácie.

Opis projektu
Cieľom projektu je priniesť nové poznatky z oblasti účinnosti algoritmov pri rozpoznávaní nevyžiadanej elektronickej komunikácie.
Zameriava sa na výskum nových algoritmov pre rozpoznávanie nevyžiadanej elektronickej komunikácie (Spam, Phishing, Scam). Skúmaním vybraných riešení budú v reálnom prostredí vyhodnotené základné princípy detekcie nevyžiadanej komunikácie a kvantitatívnymi parametrami popísaný model správania sa zdrojov, ktoré túto komunikáciu generujú.
Projekt bude orientovaný na:

- dôkladné skúmanie techník, ktoré sú využívané na maskovanie a distribúciu spamov, phishingu, scamov,
- štúdium súvislostí a charakteristík komunikácie, ktoré budú určujúce pre návrh efektívnych algoritmov na rozpoznávanie nevyžiadanej elektronickej komunikácie,
- identifikáciu slabých miest tradičných prístupov pre detekciu nevyžiadanej elektronickej komunikácie a návrh účinných algoritmov pre inteligentnú detekciu nevyžiadanej elektronickej komunikácie vo vysokorýchlostných sieťach.

Etapy:

Etapa	Popis etapy (ciele a aktivity):	Začiatok	Koniec	Počet mesiacov
1.	Skúmanie vybraných antispamových riešení, porovnanie kvality detekcie spamov, phishingu, scamov v predchádzajúcej SMTP komunikácii v reálnom prostredí.	01.08.2010	31.10.2010	3
2.	Skúmanie modelu správania sa pôvodcov spamu, metód maskovania a hromadnej distribúcie spamov	01.11.2010	31.01.2011	3
3.	Teoretický návrh efektívnych, škálovateľných algoritmov rozpoznávania nevyžiadanej elektronickej komunikácie s dôrazom na minimalizáciu výskytu false negative a false positive detekcií	01.02.2011	31.01.2012	12
4.	Návrh simulačných scenárov na verifikáciu účinnosti skúmaných algoritmov	01.02.2012	31.07.2012	6
5.	Realizácia vlastného výskumu účinnosti vyvinutých algoritmov v reálnom toku elektronickej komunikácie	01.08.2012	31.03.2013	8
6.	Vypracovanie štúdie sumarizujúcej výskumné poznatky	01.04.2013	31.07.2013	4

Financie:

Plánovaná výška oprávnených nákladov na projekt: 1979801 €
Celkom: 1979801 €
Vlastné prostriedky 0
Požadovaná dotácia 1979801 €

Položka/rok		2010	2011	2012	2013	Suma:
Bežné náklady	Priame náklady	217605	524234	523234	305728	1570801
	Nepriame náklady	4000	1000	1000	500	6500
	Spolu:	221605	525234	524234	306228	1577301
Kapitálové výdavky:		390000	12500	0	0	402500
Spolu:		611605	537734	524234	306228	1979801

Plánované výstupy riešenia:

Rok	Výstup	Publikácia (typ / počet)
2010	Teoretický model správania sa pôvodcov spamu a metód maskovania a distribúcie spamu	
2011	Návrh efektívnych algoritmov pre detekciu spamu	Teoretický model (štúdia / 1)
2012	Výskum účinnosti navrhnutých algoritmov	Návrh algoritmov (štúdia / 1)
2013	Sumarizácia poznatkov	Komparácia účinnosti algoritmov (štúdia / 1) Záverečná publikácia (publikácia / 1)

Využitie výsledkov

Výsledky projektu do značnej miery prispejú k zvýšeniu úrovne poskytovania jednej z najrozšírenejších foriem komunikácie – elektronickej pošty. Slovanet a.s. dlhodobo pôsobí ako poskytovateľ služieb internetu v celoslovenskom meradle, v súvislosti s nevyžiadanou elektronickou komunikáciou mu vznikajú permanentné náklady na hardvér, softvér a technickú podporu. Okrem toho výsledky by mali v širšom meradle eliminovať aj škody, ktoré nevyžiadaná komunikácia spôsobuje zákazníkom, nakoľko spam často figuruje ako vektor prenášajúci malware. Keďže ide o projekt základného výskumu, získané poznatky budú najskôr využité pri vývoji nového antispamového softvéru a až následne zavedené do prevádzky ako nové služby.

Vysvetlenie používaných pojmov (slovník)

V nasledujúcom texte sa vyskytujú pojmy, ktoré nie sú všeobecne známe alebo používané, preto sem zaraďujeme krátky výkladový slovníček:

bot - internetový robot - počítačový program, ktorý pre svojho majiteľa opakovane vykonáva na internete nejakú rutinnú činnosť

botnet - sieť botov

DNS - Domain Name System, poskytuje mechanizmus získania IP adresy pre každé meno systému v sieti (lookup) a naopak (reverse lookup)

fn - false negative, spam mylne klasifikovaný ako ham

fnr - false negative rate, podiel zle klasifikovaných spamov

$$= \frac{fn}{tp+fn}$$

fp - false positive, ham mylne klasifikovaný ako spam

fpr - false positive rate, podiel zle klasifikovaných hamov

$$= \frac{fp}{tn+fp}$$

ham - legitímny mail (opak spamu)

IP - Internet Protocol, dátovo orientovaný komunikačný protokol sieťovej vrstvy používaný zdrojovým a cieľovým systémom na výmenu dát prostredníctvom siete s prepínaním paketov

malware - skratka z malicious software, škodlivý software (trójske kone, vírusy, spyware, adware, atď.)

open-mail-relay - SMTP server nakonfigurovaný tak, že umožňuje odosielať maily od kohokoľvek komukoľvek

phishing - kradnutie hesiel a iných osobných údajov za účelom obohatenia sa

phisher - osoba, ktorá vykonáva phishing

proxy - server počítačovej siete, ktorý umožňuje klientom nepriame pripojenie k inému serveru; funguje ako sprostredkovateľ medzi klientom a cieľovým serverom, prekladá požiadavky klienta a oproti cieľovému serveru vystupuje ako klient. Prijatú požiadavku potom odosiela naspäť klientovi

scam - pokus o podvod, pri ktorom si podvodník najprv získa dôveru obete (v našom prípade prostredníctvom mailu)

SMTP - Simple Mail Transfer Protocol, internetový štandand pre prenos mailu prostredníctvom IP sietí

spam - nevyžiadaný, hromadne rozposielaný mail, často sa pod ním rozumie aj scam a phishing

spamer - osoba produkujúca spam

spyware - druh malware zameraný na sledovanie aktivít používateľa a ilegálne odosielanie citlivých informácií z jeho počítača

tn - true negative, správne klasifikovaný ham

tp - true positive, správne klasifikovaný spam

Hlavné realizované výstupy (výsledky) za rok 2010:

Činnosti na projekte boli v priebehu roku 2010 orientované najmä na zber informácií, testovanie niektorých antispamových produktov a tiež na prípravné práce potrebné pre ďalší výskum.

Najdôležitejšie činnosti výskumného tímu v tomto období boli:

1. zber dostupných informácií o problematike spamu a anti-spamových riešení. S ohľadom na rozsiahlosť problematiky ide o prakticky nekonečný proces, pričom špeciálnu pozornosť venujeme práve teoretickým riešeniam, ktoré ešte neboli aplikované v praxi.
2. príprava dočasného hardwarového prostredia pre potreby výskumu. Keďže obstarávacie konanie na technické prostriedky prebehne až v roku 2011, museli sme toto obdobie preklenúť náhradným riešením
3. vytvorenie testovacieho prostredia na testovanie anti-spamových riešení. Toto prostredie bolo naprogramované v jazyku Java a s využitím knižnice JavaMail, a to z dôvodu ľahkej prenositeľnosti kódu a tiež silnej celosvetovej komunitnej podpory uvedenej platformy. Ide o cca 11 000 riadkov kódu, pričom prakticky všetok kód je využiteľný aj v ďalších etapách projektu.
4. príprava mailového korpusu (množiny mailov) pre následné testovanie. Korpus pozostáva z dvoch vzoriek mailov, jedna je zo septembra a druhá z októbra 2010 a tvorí ho približne 24.000 mailov. Korpus bude následne využitý aj v rámci samotného výskumu.
5. testovanie vybraných štandardných antispamových riešení (SpamAssassin, BogoFilter) na pripravenom mailovom korpuse. Oba produkty boli testované s rôznymi testovacími scenármi. Výsledky testovania v budúcnosti použijeme v relácii k výsledkom testovania nášho vlastného navrhnutého riešenia.

Spam a jeho kategorizácia

Nevyžiadané a/alebo škodlivé maily sa nazývajú **spam** (ako opozitum k bežnej mailovej komunikácii, ktorá je nazývaná **ham**). Objem spamu sa dlhodobo (t.j. niekoľko posledných rokov) pohybuje na úrovni okolo 90% všetkej mailovej komunikácie.

Primárnym cieľom spamu je získanie finančného prospechu – či už upozornenie na produkt alebo službu, ktoré sú k dispozícii, alebo získanie osobných údajov (vrátane overených mailových adries) s cieľom následnej podvodnej manipulácie.

Dôvod existencie spamu je teda čisto finančný. Z pohľadu trhu pôsobia samotní spameri ako reklamní agenti a objednávateľom ich služieb býva väčšinou (aj legálna) obchodná spoločnosť alebo - naopak – páchatelia trestných činov.

Z pohľadu príjemcu predstavuje spam nákladovú položku – jednak z hľadiska času konkrétneho používateľa, ktorý strávi čistením svojho mail-boxu, jednak z hľadiska administrátora, na ktorého sú potom kladené zvýšené časové nároky pri zavádzaní opatrení proti spamu, a tiež z hľadiska zvýšených nákladov na komunikačnú infraštruktúru, ktorá je zaťažovaná výrazne vyšším počtom mailov. Tieto náklady sa nedajú exaktne vyčísliť, ale odhady len pre USA sú v hodnotách miliárd USD ročne.

Konkrétne číselné informácie v ďalšom texte sú prevzaté zo stránky firmy Symantec http://www.messagelabs.com/, ktorá využíva svoju celosvetovo inštalovanú sieť antispamových riešení na sumarizáciu globálnych poznatkov o aktuálnom stave šírenia spamu.

Spam z hľadiska jeho obsahu (a čiastočne aj cieľa) môžeme rozdeliť na:

1. **obchodné ponuky** s odkazom na WEB stránku – ide o najčastejší a najznámejší druh spamu. Portfólio ponúkaných produktov je veľmi široké, od lacných viagra klonov, cez rôzne finančné hry (kasína) a možnosti poistenia až po predaj elektroniky a niektorých ďalších druhov bežného spotrebného tovaru. Samotná WEB stránka môže byť využitá aj na získanie osobných údajov. Takáto stránka svojim vzhľadom napodobňuje inú – legálnu – stránku a pomýli tým používateľa, ktorý potom rutinne vyplní svoje prihlasovacie či iné osobné údaje (ide o tzv. phishing).
2. **„nigérijské listy“** – skupina mailov obsahujúca obchodnú ponuku na získanie finančného obnosu ako podielu z obchodnej operácie alebo oznam o výhre. V týchto prípadoch odosielateľ požaduje osobné údaje adresáta a/alebo priamo zaslanie finančnej čiastky ako zálohy. V prípade takýchto mailov ide o priamy pokus o podvod.
3. **kontaktné maily** – sú maily typu „*Ahoj, videla som tvoju fotku na Facebooku a chcem sa s tebou zoznámiť*“, ktorých cieľom je – v prípade, že adresát odpovie - buď vytvorenie overeného zoznamu cieľových mailových adries pre budúci spam alebo následná mailová komunikácia v duchu nigérijských listov.
4. **verifikačné maily** – ich úlohou je overiť adresu príjemcu. Takýto spam v sebe obsahuje zakódovanú identifikáciu príjemcu, a to väčšinou priamo v linke na WEB. Po kliknutí na používateľa na linku, alebo po automatickom stiahnutí obrázkov z danej WEB adresy je na strane WEB stránky zaevidované, že konkrétna emailová adresa je „živá“ a môže byť cieľom spamovej kampane. Takýmto spôsobom sa zvyšuje efektívnosť spamových kampaní.
5. **vírusy** – do tejto skupiny zahŕňame všetky maily, ktoré v sebe obsahujú škodlivý kód (priložený .exe súbor, kód pre screensaver, javascripty, atď.) alebo linku na WEB stránku s takýmto kódom. Názov „vírus“ je v tomto prípade chápaný trochu univerzálnejšie, zahŕňa v sebe jednak klasické vírusy, ale tiež kód pre sledovanie aktivít používateľa (získavanie osobných údajov a ich odosielanie) a kód pre generovanie spamu.

Šírenie a maskovanie spamu

Ako už bolo spomenuté vyššie, približne 90% všetkej mailovej komunikácie tvorí spam. Ide o spôsob reklamy, ktorý je pre zadávateľa relatívne lacný, a to aj napriek svojej nízkej efektívnosti. Paul Graham na svojej stránke http://www.paulgraham.com/spam.html udáva, že z 1 milióna príjemcov spamu len 18 adresátov na spam pozitívne zareaguje. Inými slovami, správanie 0,0018% adresátov spamu je dôvodom na vygenerovanie 90% mailovej komunikácie. Iné materiály uvádzajú ešte nižšiu účinnosť spamu.

Pritom sa udáva, že za viac než 80% spamu, ktorý je rozposlaný v rámci Európy a Severnej Ameriky, je zodpovedných menej než 200 spamerov. Časť z nich je identifikovaná v tomto zozname: http://www.spamhaus.org/rokso/index.lasso.

Existujú dva základné spôsoby odosielania spamu:

1. maily sú odosielané priamo z IT infraštruktúry pôvodcu spamu
2. maily sú odosielané zo zavírených počítačov používateľov, ktorí ani netušia, že sú producentami spamu

Prvý spôsob je zastúpený len obmedzene, pretože aj pri maskovaní odosielateľa dochádza k tomu, že jeho IP adresa je skôr či neskôr odhalená. Potom sa dostáva na tzv. black-list a na úrovni internet providera alebo mail servera je následne možné blokovať akúkoľvek mailovú komunikáciu z danej IP adresy. Tým sa pochopiteľne efektívnosť spamu značne znižuje.

Oveľa častejší je preto druhý spôsob. Ten vyžaduje, aby bol najprv infikovaný počítač náhodného používateľa škodlivým kódom (mohol mu byť zaslaný mailom alebo si ho stiahol z WEB stránky). Takýto kód sa nazýva **bot** (od slova „robot“) a skupina (sieť) takto postihnutých počítačov **botnet**. Jednotliví bot-i sú prostredníctvom internetu riadení zo svojej centrály, odkiaľ dostávajú pokyny týkajúce sa adresátov a obsahu odosielaných mailov. V súčasnosti je prostredníctvom botnetov odosielaných až 95% všetkých spamov. Kadencia odosielania mailov je pritom pomerne vysoká, jeden bot odosiela desiatky až stovky mailov za minútu.

V oboch prípadoch sa na šírenie spamu používa tzv. **open-mail-relay**. Ide o zle zabezpečené SMTP servery, ktoré prijímajú a posielajú ďalej nielen maily od svojich registrovaných používateľov, ale od kohokoľvek z internetu. Ďalšou možnosťou pre spamerov je zriadenie SMTP servera ako súčasť bot-a, ale táto nie je až tak využívaná vzhľadom na to, že väčšina infikovaných počítačov má dynamicky prideľované IP adresy a tak vzhľadom na zapnutý „graylisting“ u cieľových serverov je úspešnosť doručenia spamu značne redukovaná.

Odosielané spamy sú len málokedy posielané ako otvorený spam. Väčšina spamerov má snahu zamaskovať skutočného odosielateľa a častokrát aj obsah mailu, aby nebol pre anti-spamové nástroje príliš ľahko identifikovateľný.

Najčastejšie ide o tieto spôsoby maskovania:

1. modifikácia hlavičiek (headers) odosielaných mailov – cieľom je vytvoriť dojem, že odosielateľom mailu je niekto iný a tak zamaskovať jeho skutočný pôvod. Vzhľadom na nedostatočne ošetrené aspekty autorizácie a autentifikácie odosielateľa mailu v SMTP protokole ide o veľmi úspešný spôsob maskovania pôvodu spamu.
2. modifikáciu textu správy – aby spamer obišiel antispamovú ochranu, modifikuje niektoré kľúčové slová (viagra -> viiagra, buy -> b u y, price -> pr1ce) alebo v prípade HTML mailu povkladá do slova písmená s fontom s veľkosťou 0 (ktoré sa používateľovi nezobrazia). Spam odosielaný v rámci tej istej kampane môže mať pri každom odoslaní čiastočne pozmenený obsah – mení sa jeho dĺžka, pridávajú sa (niekedy nezmyselné) slová na zmätenie antispamových nástrojov a pod.

3. skrývanie textu – samotný mail je na pohľad „nevinný“ (obsahuje napr. citát z nejakej knihy) a vlastné telo spamu je uložené v prílohe (PDF súbor, obrázok, ...). Príloha v sebe opäť obsahuje linku na WEB stránku

Obrana proti spamu

Na obranu proti spamu bolo vyvinutých množstvo nástrojov. Prakticky všetky implementujú jednu alebo viaceré z nasledujúcich techník:

1. analýza odosielateľa
2. odhaľovanie nekorektných záznamov v hlavičkách mailu
3. používanie vybraných kľúčových slov
4. vyhľadávanie vybraných konštrukcií v tele mailu (napr. text písaný bold fontom, červenou farbou a končiaci výkričníkom)
5. používanie štatistických metód na vyhodnocovanie tela mailu a určovanie pravdepodobnosti, že mail je spam

V ďalšom uvádzame prehľad najznámejších používaných techník obrany proti spamu:

1. Black-listing je technika založená na vytváraní zoznamov (IP adries) producentov spamu. Mail server overuje IP adresu, z ktorej prijíma mail voči jednému alebo viacerým takýmto zoznamom a pokiaľ je odosielateľ na black-liste, je mail odmietnutý.
2. White-listing je technika zoznamov s presne opačným významom, mail od odosielateľa registrovaného v zozname je pre príjemcu chápaný ako legálny.
3. Grey-listing je založený na vlastnosti mail serverov, ktoré pri neúspešnom doručení zaradia správu do fronty a pokúsia sa ju po určitom čase znovu doručiť, o čo sa spameri väčšinou nesnažia. Príjemca preto prvýkrát prijatý mail odmietne s dočasnou chybou a až pokiaľ je po čase doručený opäť, je ochotný ho akceptovať.
4. Challenge-response je metóda výzvy a odpovede. Ide o vynútenú kontrolu identity pri prijímaní správy od neznámeho odosielateľa. Správa je mail serverom ešte pred doručením podržaná vo fronte a odosielateľovi je zaslaná žiadosť o autorizáciu. Pokiaľ odosielateľ tejto žiadosti vyhovie, je správa doručená príjemcovi.
5. Kontrola legitimity doručovanej správy má snahu verifikovať odosielateľa správy (ochrana proti bot-om). Existuje niekoľko základných riešení:
   - Reverse DNS Lookup – mailserver, ktorý prijíma správu, overuje, či IP adresa odosielajúceho mailserveru zodpovedá doméne, ktorá je uvedená v poli *from*
   - Sender Policy Framework (SPF) zavádza do DNS záznamov informáciu o serveroch, ktoré sú oprávnené odosielať správy z danej domény.
   - Sender ID je veľmi podobná technika od Microsoftu
   - DomainKeys od Yahoo pracuje na princípe elektronického podpisu – odosielajúci server do odchádzajúceho mailu doplní podpis kryptovaný privátnym kľúčom, pričom pravosť tohto podpisu je kýmkoľvek verifikovateľná za pomoci verejného kľúča publikovaného v DNS

6. Kontrola správnosti hlavičiek mailu – využíva sa validácia hlavičiek vzhľadom na špecifikáciu SMTP protokolu a vzájomný vzťah údajov uvádzaných v hlavičkách mailu.
7. Slovné filtre – najjednoduchší, ale pomerne efektívny druh obrany. Filtruje maily, ktoré obsahujú slová a frázy zo zoznamu „zakázaných" slov.
8. Rule-based Scoring Systems (skórovacie systémy založené na pravidlách) – tiež kontrolujú výskyt fráz, ale na ich hodnotenie používajú sofistikovanejšie pravidlá
9. Obsahová analýza mailu (Bayes filter) je založená na využití štatistickej metódy vyhodnocujúcej pravdepodobnosť, ako často sa konkrétne slovo alebo skupiny slov daného mailu vyskytujú v spamoch.

Pravdepodobne najpopulárnejší anti-spamový nástroj je open-source produkt SpamAssassin. Tento zhodou okolností využíva všetky vyššie uvedené techniky týkajúce sa black-listov a analýzy hlavičiek a tela mailu.

Problémy obrany proti spamu

Účinnosť každého anti-spamového nástroja sa udáva v dvoch hodnotách:

1. aké percento ham-ov vyhodnotil chybne ako spam (v anglickej literatúre sa toto nazýva false-positive)
2. a aké percento spamov vyhodnotil ako ham – teda ich nezachytil (false-negative)

Príjemca mailu je vysoko citlivý práve na false-positive, pretože to znamená, že očakávaný „zdravý" mail preňho je buď označený ako spam a zaradený do príslušného mail foldra alebo mu vôbec nie je doručený.

False-negative je z pohľadu príjemcu len nepríjemné, pretože sa mu mailová schránka zapĺňa spamom, ale nedochádza k strate informácie.

Základným problémom pri identifikácii spamu je, že čím prísnejšie parametre si zadefinujeme pre jeden parameter, tým citlivejšie reaguje druhý. Anti-spamové nástroje balansujú medzi týmito dvoma protichodnými parametrami, ale prakticky u všetkých sa prejavuje efekt, že čím lepšie je vyladená ich konfigurácia pre zachytenie spamu, tým vyššie je riziko false-positive.
Ďalšie problémy súvisia s tým, že spameri sú logicky vždy o krok vpredu pred anti-spamovými nástrojmi. To sa týka jednak vyhodnocovania formálnych a obsahových náležitosti mailu, ale aj použitia štatistických metód na kvalifikovanie textového obsahu tela mailu.
V prvom prípade sa anti-spamový nástroj zameriava vyslovene na odhaľovanie chýb spamera (formálna chyba v hlavičke mailu v rozpore so špecifikáciou, vzájomný nesúlad viacerých hlavičiek, používanie už identifikovaných textových konštrukcií). Pre spamera je pomerne jednoduché svoju chybu rýchlo napraviť: napr. v súvislosti so SpamAssassinom sa udáva rýchlosť reakcie spamera – a teda strata účinnosti konkrétneho pravidla antispamového riešenia – na dni až týždne.
V druhom prípade, pri použití štatistických metód, je anti-spamový nástroj silne závislý od reakcie konkrétneho používateľa, ktorého chráni. Pokiaľ používateľ aktívne neposkytuje anti-spamovému nástroju informácie o jeho chybnom oklasifikovaní mailu, tento nemá možnosť

sa učiť a jeho účinnosť rapídne klesá. Navyše, spameri majú snahu štatistické prístupy zmiasť používaním čiastočne odlišnej slovnej zásoby a obaľovaním samotnej spamovej správy množstvom slovného balastu.
Samostatným problémom je orientácia producentov anti-spamových riešení na anglicky hovoriace prostredie. Slovenčina so svojou bohatou sémantikou predstavuje pre anti-spamové nástroje veľkú výzvu. A keďže spamy v slovenčine sa už začínajú šíriť (aj keď zatiaľ vo všetkých nami identifikovaných prípadoch ide o automatizované preklady anglických spamov), z hľadiska obsahovej analýzy plánujeme zamerať náš výskum aj týmto smerom.

Merania účinnosti anti-spamových riešení

V priebehu roku 2010 sme vykonali aj merania účinnosti dvoch voľne dostupných antispamových nástrojov: SpamAssassin 3.3.1 a BogoFilter 1.2.2.
Merania boli prevedené na oboch produktoch s default nastavením (definovaným dodávateľom), pričom boli testované na dvoch vzorkách (korpusoch) mailov v celkovom počte cca. 24 000 získaných v septembri a októbri 2010, s podielom spamu 12,5% a 21,5%. Testované boli všetky zabudované pravidlá SpamAssassina a bayesovské filtre oboch produktov, a to na jednotlivých vzorkách, s rôznymi scenármi trénovania filtrov a aj v závislosti od jazyka, v ktorom bol mail napísaný.
Najpodstatnejšie výsledky sú zhrnuté v nasledujúcej tabuľke. Tá udáva chybu *fnr* (aké % spamov zo všetkých spamov je omylom vyhodnotené ako ham, t.j. ako legitímny mail), ak požadujeme *fpr = 1%* (t.j. ak len 1 ham zo 100 môže byť omylom vyhodnotený ako spam).

Vzorka	SA: Fix + Bayes	SA: Fix	SA: Bayes	BF
September - komplet	29%	75%	10%	11%
Október - komplet	73%	91%	19%	18%
Okt/Sep - komplet			78%	14%
September - anglicky	29%	69%	30%	23%
Október - anglicky	89%	88%	100%	100%
Okt/Sep - anglicky			87%	77%
September - slovensky	60%	86%	51%	53%
Október - slovensky	40%	45%	34%	32%
Okt/Sep - slovensky			47%	28%

Z výsledkov vyplýva, že zabudované pravidlá SpamAssassina (v tabuľke sú označené ako *Fix*), podobne ako aj bayesovské filtre oboch produktov, majú pri požiadavke 1% fpr pomerne veľkú chybovosť a prepúšťajú neúmerne veľa spamu.

Svoju úlohu v tomto prípade zohráva samotný testovací korpus, ktorý pomerne verne odráža skladbu mailov v prostredí slovenského internetu, ale aj konfigurácia oboch produktov. Kým v prípade bayesovských filtrov sa lepšie hodnoty dajú získať len veľmi obtiažne, pravidlá (resp. ich váhy = vplyv na klasifikáciu mailu) SpamAssassina je možné optimalizovať, a ako ukázali naše výpočty, môže sa tak dosiahnuť výrazné zlepšenie výsledkov.

Popis prínosov za rok 2010:

Hlavným prínosom za rok 2010 je vybudovanie mailového korpusu určeného na testovanie úspešnosti detekcie spamu a tiež implementácia príslušného testovacieho prostredia (frameworku) ako nutnej podmienky pre ďalší výskum.

Zároveň boli výskumným tímom sumarizované informácie o produkcii a distribúcii spamu, ako aj o používaných a skúmaných metódach detekcie spamu. Tieto budú predmetom samostatnej publikácie - štúdie v roku 2011.

Z testov SpamAssassina ako vybraného referenčného antispamového riešenia voči nášmu mailovému korpusu vyplýva možnosť štatistickej optimalizácie ohodnotenia (skóre) parametrov tohto nástroja.

Zároveň tieto testy ukázali, že v reálnom prostredí slovenského internetu je účinnosť testovaných anti-spamových nástrojov výrazne nižšia, než aká je dosahovaná na voľne dostupných korpusoch, na ktoré sú tieto nástroje pravdepodobne aj optimalizované.

Hlavné realizované výstupy (výsledky) za rok 2011:

Koncom januára 2011 bola zverejnená plánovaná štúdia „*Problematika spamu, jeho pôvodu a distribúcie a výskum efektívnosti vybraných antispamových riešení*“. Štúdia je voľne dostupná na internete na adrese http://spamia.slovanet.sk/StudiaSpamia2011.pdf. Táto štúdia sumarizuje poznatky zo skúmanej oblasti, ktoré výskumný tím zhromaždil v priebehu roka 2010. Okrem toho však obsahuje aj rozsiahle merania účinnosti voľne dostupných antispamových produktov SpamAssassin 3.3.1 a BogoFilter 1.2.2.

Merania účinnosti vybraných komerčných antispamových riešení

V roku 2011 bola publikovaná štúdia *„Účinnosť vybraných komerčných antispamových produktov“* sumarizujúca účinnosť vybraných komerčných produktov FortiMail a FortiGate. Štúdia je voľne dostupná na adrese http://spamia.slovanet.sk/komercny-antispam-studia.pdf. Na testovanie komerčných antispamových nástrojov boli celkovo použité tri korpusy (označené Sep, Okt, Mar) s celkovým počtom cca. 40 tisíc emailov. Prevažná väčšina z nich bola v slovenčine resp. češtine, s výrazným zastúpením anglických a nemeckých emailov. Ostatné jazyky boli reprezentované menším počtom emailov. Podiel spamu v korpusoch bol v rozmedzí 11 – 30%.

Nasledujúca tabuľka zobrazuje výsledky testovania komerčných antispamových produktov:

Produkt / filter resp. podmienka	fpr (%)	fnr (%)
FortiGate 4.0	2,3	68,0
FortiMail 4.0 - FortiGuard	3,6	40,0
FortiMail 4.0 – Bayes (Okt/Sep)	0,1	15,6
FortiMail 4.0 – Bayes (Mar/Sep)	0,03	2,8

Aj tieto merania potvrdili pomerne nízku úspešnosť detekcie spamu, pokiaľ je založená na heuristike (rôznou formou implementované pravidlá).
Naopak, veľmi dobré výsledky dosiahol bayesovský filter zabudovaný vo FortiMaile, zvlášť na našom poslednom (najnovšom) korpuse. Potvrdilo sa teda, že použitie štatistickej metódy (v tomto prípade Bayesovho filtra) je pri vhodnom trénovaní efektívnejšie než používanie fixných pravidiel.

Teoretický návrh algoritmu na filtrovanie spamu

V rámci roku 2011 bola rozhodujúca časť úsilia výskumno-vývojového tímu venovaná plneniu 3. etapy projektu, teda teoretickému návrhu efektívneho a škálovateľného algoritmu na rozpoznávanie nevyžiadanej elektronickej komunikácie s dôrazom na minimalizáciu výskytov false negative a false positive detekcií.

Na základe získaného prehľadu existujúcich riešení na filtrovanie elektronickej komunikácie, získaného počas 1. a 2. etapy projektu, je možné tvrdiť, že ich podstatnou zložkou je súbor heuristických pravidiel a textová a lexikálna analýza s využitím postupov a algoritmov text-miningu, akými sú napríklad štandardizácia dokumentu, tokenizácia, lematizácia (vyhľadávanie koreňov slov), určovanie slovných druhov, parsovanie, vyhľadávanie kľúčových slov a iné. Súčasťou filtrov sú aj postupy využívajúce globálne a lokálne informácie o emailovej komunikácii, napr. blacklisty či whitelisty.
Spomedzi nevýhod existujúcich riešení môžeme spomenúť vysokú vulnerabilitu (zraniteľnosť) algoritmov, závislosť na jazyku, fixovanosť heuristických pravidiel, viazanosť na binárnu klasifikáciu a rigidnosť výsledného hodnotenia spamovosti emailu. Preto sme sa snažili o odlišné prístupy k filtrovaniu a kategorizácii emailov.

Na základe uvedených poznatkov sme navrhli filtrovanie a kategorizáciu emailov založenú na kvantitatívnych profiloch. Namiesto reprezentácie emailu pomocou množiny slov/tokenov, email reprezentujeme vektorom reálnych/celých čísel vopred zvolenej dimenzie; takýto vektor nazývame kvantitatívny profil emailu.
Kvantitatívne profily zachytávajú charakteristiky emailov, ktoré doposiaľ neboli využívané v existujúcich spamových filtroch, ani vedecky skúmané v odbornej literatúre (jediná nám známa výnimka je článok *„Email shape analysis“* autorov Sroufe, Phithakkitnukoon, Dantu a Cangussu z roku 2010). V práci navrhujeme niekoľko rôznych profilov, ktoré možno rozčleniť na dva typy: elementárne a symbolicko-dynamické, tzv. RQA-profily. Jeden z elementárnych kvantitatívnych profilov je riadkový profil, tvorený dĺžkami riadkov. Riadkový profil je špeciálnym prípadom všeobecného binárneho profilu. Binárny profil si všíma „doby“ medzi opakovanými výskytmi konkrétneho patternu v emaili. V prípade riadkového profilu je patternom znak konca riadku. Ďalším jednoduchým profilom je znakový

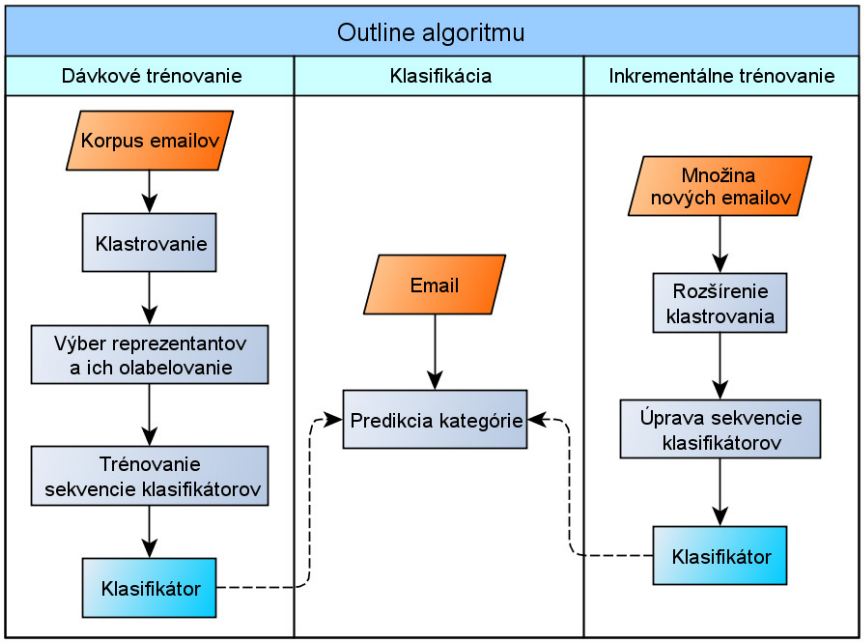
profil, ktorý je v základnej forme vlastne histogramom početností výskytov jednotlivých znakov, respektíve skupín znakov. Zložitejším typom kvantitatívnych profilov sú RQA-profily. Ide o profily založené na novom symbolicko-dynamickom prístupe k emailu; voľne povedané, email chápeme ako realizáciu stacionárneho stochastického informačného zdroja a RQA-profil vystihuje charakteristiky tohto zdroja. Takou charakteristikou je napríklad determinizmus, zachytávajúci „predpovedateľnosť“ informačného zdroja. Ďalšou je Shannonova entropia.

Jednou z prehliadaných oblastí v problematike spamu je labelovanie a relabelovanie korpusu emailov. Podľa nášho názoru však ide o vec zásadnej dôležitosti, nakoľko spätná väzba od používateľa nie je vždy dostupná v potrebnej miere. V takom prípade je nesmierne dôležité minimalizovať náklady s tým spojené. Z tohto dôvodu je neoddeliteľnou súčasťou algoritmu adaptívne klastrovanie. Okrem efektívneho výberu vzorky pre labelovanie, adaptívne klastrovanie, vďaka stanoveniu váh olabelovaných emailov, umožňuje aj redukciu trénovacej vzorky potrebnej pre klasifikovanie.

Z dôvodu efektívnosti sa v algoritme samotné klasifikovanie vykonáva sekvenčne, pričom sa postupne mení sada kvantitatívnych profilov použitých ako prediktory. Začína sa s jednoduchšie získateľnými profilmi. Ako klasifikačný algoritmus je vhodné použiť Random Forest; medzi jeho silné stránky patrí robustnosť vzhľadom na extrémne profily, invariantnosť na monotónne transformácie profilov, jednoduchá paralelizovateľnosť a hlavne vynikajúca výkonnosť.

Časovo najnáročnejšou fázou každého učiaceho sa filtra je jeho trénovanie. Keďže spam sa v čase dynamicky mení, je potrebné filter opakovane pretrénovávať. Za týmto účelom súčasťou aj tohto algoritmu je takzvané inkrementálne trénovanie. Jeho úlohou je, na základe nových emailov korpusu, inovovať klastrovú štruktúru, váhy emailov, ako aj samotnú sekvenciu klasifikátorov s pomerne nízkymi výpočtovými nárokmi.

Globálny pohľad na funkčné celky algoritmu podáva nasledujúci diagram:



Detailný popis algoritmu je obsahom štúdie *„Teoretický návrh algoritmu na filtrovanie spamu“* http://spamia.slovanet.sk/navrh-algoritmu.pdf. Jeho nosnými prvkami sú

- kvantitatívne profily a
- adaptívne klastrovanie.

Každá z navrhovaných inovácií prináša viacero výhod predkladaného algoritmu. Medzi výhody, vyplývajúce z reprezentácie emailov kvantitatívnymi profilmi, patria hlavne:

- škálovateľnosť, primeraná výpočtová zložitosť, paralelizovateľnosť, robustnosť,
- nízka miera vulnerability,
- kombinovateľnosť s existujúcimi riešeniami,
- adaptívnosť klasifikovania emailu,
- flexibilnosť a rozšíriteľnosť algoritmu,
- možnosť zatrieďovania emailov do jemnejších skupín,
- nezávislosť na jazyku.

Vďaka tomu, že klasifikovanie/kategorizácia emailov je založená na Random Foreste, je výsledný algoritmus paralelizovateľný a vysoko škálovateľný, ako aj robustný voči extrémnym emailom. Vysoká odolnosť algoritmu voči prelomeniu spamermi (nízka miera vulnerability) je zabezpečená variabilnosťou reprezentácie emailu pomocou kvantitatívnych profilov a tým, že výber profilov použitých na klasifikovanie nie je dopredu stanovený, ale závisí od konkrétneho korpusu, prípadne od používateľových preferencií. Keďže email je reprezentovaný kvantitatívnym profilom a nie množinou slov/tokenov, algoritmus extrahuje z emailu inú informáciu než štandardné filtre; táto ortogonalita umožňuje kombinovanie algoritmu s existujúcimi filtrami a tým dosiahnutie vyššej efektívnosti. Emaily sa kategorizujú pomocou postupnosti klasifikátorov, pričom sa začína od klasifikovania pomocou jednoduchých kvantitatívnych profilov a, ak je to potrebné, postupuje sa k čoraz zložitejším profilom. Nosné prvky algoritmu sú konfigurovateľné, t.j. kvantitatívne profily a klastrovacie metriky možno zvoliť v závislosti od korpusu; takisto je možné algoritmus jednoducho rozšíriť definovaním vlastného kvantitatívneho profilu respektíve metriky. Veľakrát nepostačuje hrubé členenie emailov na spam a ham, ale je potrebné emaily zatrieďovať do jemnejších kategórií; predkladaný algoritmus umožňuje kategorizovať emaily do používateľom zadaných skupín. Na rozdiel od text-miningového prístupu k filtrovaniu emailov, ktorý je závislý od gramatických a lexikálnych pravidiel konkrétneho jazyka, náš prístup pomocou kvantitatívnych profilov je jazykovo nezávislý.

Vďaka ďalšej inovácii založenej na adaptívnom klastrovaní, predkladaný algoritmus získava nasledovné vlastnosti:

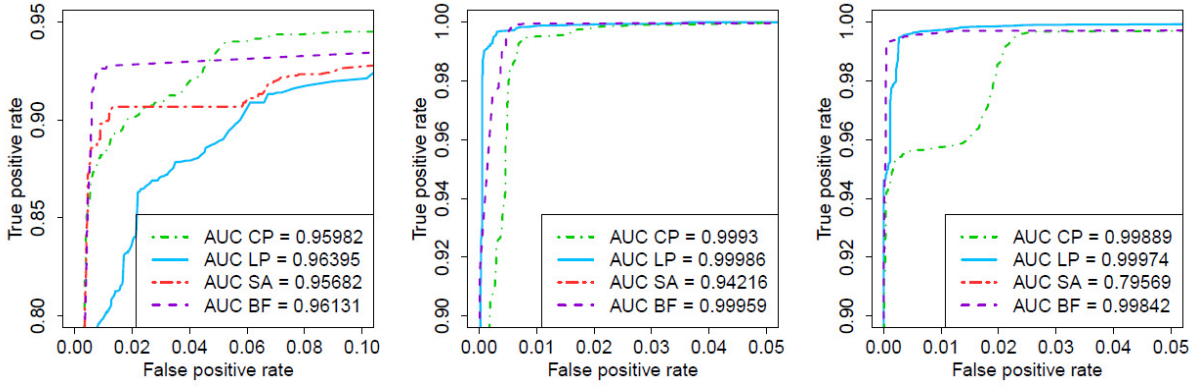
- zohľadnenie informácie aj z nelabelovaných emailov,
- efektívne využívanie spätnej väzby o labeli emailu,
- rýchlejšie rozpoznanie novej spamovej kampane,
- selektívny výber trénovacej množiny emailov.

Keďže vstupom do algoritmu je celý korpus emailov (olabelovaných aj neolabelovaných), informácia z neolabelovaných emailov podstatným spôsobom ovplyvňuje klastrovaciu štruktúru aj samotné predikovanie kategórie emailov. V prípade získania labelu emailu od používateľa je táto informácia automaticky rozšírená na celé „okolie“ tohto emailu. Spamy z novej kampane sa prejavia ako nový klaster, čo má následne vplyv na klasifikovanie. Na

manuálne labelovanie sú vybrané len tie emaily, ktoré reprezentujú klastre získané v adaptívnom klastrovaní, pričom homogénne klastre sú zastúpené nižším počtom reprezentantov ako nehomogénne.

V rámci predbežného skúmania reprezentácie emailov kvantitatívnymi profilmi bol publikovaný odborný článok *„Spam filtering by quantitative profiles“* http://arxiv.org/abs/1201.0040. V článku je porovnaná úspešnosť klasifikovania emailov pomocou dvoch základných kvantitatívnych profilov (riadkový a znakový profil) a voľne dostupných antispamových riešení SpamAssassin a Bogofilter. Dva kvantitatívne profily s využitím Random Forest algoritmu dosahujú prinajmenšom porovnateľné výsledky ako optimálne nakonfigurované heuristické pravidlá SpamAssassina a dávkovo trénovaný Bogofilter. Pritom získanie kvantitatívnych profilov emailov je implementačne jednoduché, zároveň výpočtovo a pamäťovo nenáročné, čo je ukázané v doplňujúcej štúdii *„Supplementary material to Spam filtering by quantitative profiles“* http://www.savbb.sk/~grendar/spam/SupplementToQuantitativeProfiles.pdf.
Porovnanie bolo realizované na privátnom korpuse a dvoch verejných korpusoch TREC07 a CEAS08. Výsledky riadkových a znakových profilov sú sumarizované v tabuľke a na grafoch ROC kriviek. Na privátnom korpuse dosahujú najlepšiu úspešnosť SpamAssassin (SA) a Bogofilter (BF), pričom znakový profil (CP) je len mierne horší. Najmenej efektívnym je riadkový profil (LP).
Na verejných korpusoch dosahujú všetky filtre s výnimkou SpamAssassina (SA) výrazne lepšiu výkonnosť ako na privátnom korpuse. Oba nami navrhované kvantitatívne profily (CP, LP) majú ďaleko vyššiu efektívnosť ako SpamAssassin (SA). Pri hodnote fpr = 0.5% je riadkový profil dokonca úspešnejší ako Bogofilter (BF). Poznamenajme však, že efektívnosť filtrov na verejných korpusoch je príliš optimistická z dôvodu jednoduchej odlíšiteľnosti hamov a spamov na základe hlavičiek emailov, viď článok *„Spam filtering by quantitative profiles“* http://arxiv.org/abs/1201.0040.

filter	privátny (Okt/Sep)		TREC07		CEAS08	
	pri 0.5%	pri 1%	pri 0.5%	pri 1%	pri 0.5%	pri 1%
CP	14.39	11.64	2.53	0.49	4.38	4.25
LP	21.33	20.10	0.30	0.13	0.39	0.27
SA	12.68	10.10	35.87	30.51	76.14	69.92
BF	13.05	7.38	0.40	0.06	0.47	0.36



Na základe spomínanej štúdie výkonnosti dvoch základných kvantitatívnych profilov je možné očakávať, že navrhovaný algoritmus bude dosahovať porovnateľnú, ak nie lepšiu výkonnosť než existujúce voľne dostupné, respektíve komerčné riešenia na filtrovanie spamu. Takisto predbežné výskumy účinnosti adaptívneho klastrovania ukazujú, že by malo byť možné týmto spôsobom znížiť náročnosť labelovania a relabelovania korpusu, ako aj zvýšiť výkonnosť klasifikácie. Tieto vlastnosti spolu s ďalšími výhodami vrátane škálovateľnosti, flexibility a rozšíriteľnosti robia navrhovaný algoritmus zaujímavým pre využitie v reálnej internetovej komunikácii.

Popis výstupov za rok 2011:

Plánovaným prínosom za rok 2011 je publikovanie štúdií „*Problematika spamu, jeho pôvodu a distribúcie a výskum efektívnosti vybraných antispamových riešení*“ http://spamia.slovanet.sk/StudiaSpamia2011.pdf a *„Účinnosť vybraných komerčných antispamových produktov“* http://spamia.slovanet.sk/komercny-antispam-studia.pdf.

Hlavným prínosom je teoretický návrh algoritmu sumarizovaný v štúdii *„Teoretický návrh algoritmu na filtrovanie spamu“* http://spamia.slovanet.sk/navrh-algoritmu.pdf. Okrem toho bol publikovaný odborný článok *„Spam filtering by quantitative profiles“* http://arxiv.org/abs/1201.0040 spolu s doplňujúcim materiálom *„Supplementary material to Spam filtering by quantitative profiles“*, ktorý je verejne dostupný na adrese http://www.savbb.sk/~grendar/spam/SupplementToQuantitativeProfiles.pdf.

Hlavné realizované výstupy (výsledky) za rok 2012:

Návrh nových kvantitatívnych profilov

Bolo zavedených viacero nových tried kvantitatívnych profilov na reprezentáciu emailov, ktoré podstatne rozšírili paletu profilov popísaných v článku „Spam filtering by quantitative profiles“. Okrem nových inštancií binárnych profilov (slovný, vetný, zátvorkový a iné) boli zavedené nasledovné triedy kvantitatívnych profilov:

- zoskupené znakové profily,
- d-gramové zoskupené znakové profily,
- „moving-window“ profily,
- histogramové binárne profily a
- veľkostné profily.

V daných triedach bolo skonštruovaných niekoľko desiatok inštancií nových kvantitatívnych profilov. Výkonnosť klasifikovania emailov bola posúdená na privátnych, ako aj na dvoch najčastejšie analyzovaných verejných korpusoch TREC a CEAS.

Návrh tried simulačných scenárov

Na základe poznatkov o aktívne sa brániacich filtroch a multi-klasifikačných systémoch, ako aj štandardných metodikách používaných pri verifikovaní účinnosti anti-spamových riešení, bolo navrhnutých viacero tried simulačných scenárov na verifikáciu účinnosti algoritmu SPAMIA.

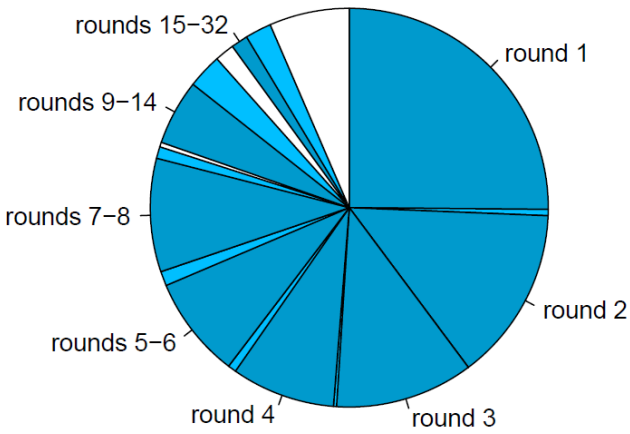
Klastrovací algoritmus AD-OPTICS

Bol navrhnutý a zaimplementovaný adaptívny dynamický klastrovací algoritmus AD-OPTICS, ktorý je jedným z hlavných pilierov algoritmu SPAMIA. Je založený na metóde OPTICS/DBSCAN, ktorá je využívaná opakovane. Obrovskou výhodou klastrovania pomocou AD-OPTICS je jeho jednoduchá konfigurovateľnosť. Nutnosť zložitej, často manuálnej voľby maximálneho klastrovacieho polomeru pre OPTICS, bola nahradená automatizovanou voľbou založenou na kvantiloch vzdialeností *k*-tych najbližších susedov a odhadu času na získanie príslušných okolí klastrovaných objektov. Na kvantiloch založený prístup bol využitý aj pri extrakcii klastrov z výstupu algoritmu OPTICS. Po analýze rôznych metód popísaných v odbornej literatúre bol využitý prístup, popísaný v štúdii o návrhu algoritmu; konkrétne boli využité tzv. flat-DBSCAN klastre, pričom na výber konkrétnych klastrovacích polomerov boli využité kvantily reachability-distance, hlavného výstupu OPTICS-u.

Pre overenie prístupu a posúdenie efektívnosti bol AD-OPTICS použitý na rozklastrovanie verejne dostupných korpusov TREC07 a CEAS08, obsahujúce približne 75 tisíc a 138 tisíc emailov. Nastavenie bolo zvolené nasledovne: minimálne 100 emailov v klastri, maximálne 20 minút na jeden beh OPTICS-u, ukončenie pri menej ako 5000 objektoch. Výsledok je popísaný v nasledovnej tabuľke. Z nej vidíme, že prevažná väčšina emailov (72% z TREC07 a 83% z CEAS08) bola zaradená do úplne homogénnych klastrov, tj. do klastrov, obsahujúcich iba spamy alebo iba hamy.

Korpus	klastrov	100%	90-100%	0-90%	čistota
TREC07	431	331 (72%)	58 (13%)	42 (15%)	94,3%
CEAS08	724	638 (83%)	55 (8,5%)	31 (8,5%)	97,5%

Na obrázku je znázornený samotný priebeh klastrovania vzorky CEAS08 podľa jednotlivých behov algoritmu. Vidíme, že v prvých 8 behoch boli takmer všetky klastre úplne homogénne. V neskorších behoch už vznikali aj nehomogénne klastre, avšak aj ku koncu algoritmus stále nachádzal 100% čisté klastre.



Tieto výsledky poukazujú na výhodnosť využitia klastrovania ako nosného prvku filtrovacieho algoritmu SPAMIA. Návrh klastrovacieho algoritmu AD-OPTICS bol spracovaný vo forme článku „OPTICS-based clustering of emails represented by quantitative profiles“.

Popis výstupov za rok 2012:

Výstupom z riešenia tretej etapy projektu (obdobie 1.2.2011 – 31.1.2012) bola štúdia „Teoretický návrh algoritmu na filtrovanie spamu“, publikovaná na http://spamia.slovanet.sk/navrh-algoritmu.pdf. Táto štúdia popisuje návrh algoritmu umožňujúceho efektívnejšiu detekciu nevyžiadanej elektronickej komunikácie (spamu).

V roku 2012 boli publikované nasledovné odborné články a prezentácie:

- „Spam filtering by quantitative profiles“, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012
- „Email categorization and spam filtering by Random Forest with new classes of quantitative profiles“, COMPSTAT 2012, A. Colubi et al., The International Statistical Institute/International Association for Statistical Computing, 2012
- „Supplement to Email categorization and spam filtering by Random Forest with new classes of quantitative profiles“
- „Spam a jeho detekcia“, Informačná bezpečnosť 2012, sasib, 2012
- „SPAMIA: Spam Filtering by Quantitative Profiles “, Applied Statistics 2012, Statistical Society of Slovenia, 2012
- „SPAMIA: filtrovanie spamu pomocou kvantitatívnych profilov a adaptívneho klastrovania”, Bez(a)Dis, Ústav informatiky PF UPJŠ Košice, 2012

Hlavné realizované výstupy (výsledky) za rok 2013:

Rekurenčná kvantifikačná analýza

Vo februári 2013 bol na preprintovom serveri arxiv.org publikovaný článok „Strong laws for recurrence quantification analysis“, obsahujúci formuláciu a dôkaz silných zákonov veľkých čísel pre hlavné RQA charakteristiky. Tieto výsledky umožňujú lepšie pochopiť rekurenčnú analýzu a jej potenciál na využitie pri filtrovaní spamu. Článok bol prijatý na publikovanie v karentovanom vedeckom časopise International Journal of Bifurcation and Chaos.

Komparatívna štúdia účinnosti algoritmu Spamia

Piata etapa, venovaná realizácii vlastného výskumu účinnosti vyvinutých algoritmov v reálnom toku elektronickej komunikácie, bola zavŕšená publikovaním štúdie „Komparatívny výskum účinnosti algoritmu Spamia“ http://spamia.slovanet.sk/vyskum-ucinnosti.pdf. Obsahom štúdie je popis výsledkov testovania algoritmu Spamia, vybraných voľne dostupných antispamových produktov (SpamAssassin a Bogofilter) a komerčného produktu FortiMail.

AD-OPTICS

Jeden z nosných pilierov algoritmu Spamia – adaptívny klastrovací algoritmus AD-OPTICS – bol popísaný v článku „OPTICS-based clustering of emails represented by quantitative profiles“ a prezentovaný na odbornej konferencii Distributed Computing and Artificial Intelligence 2013, ktorá sa konala v dňoch 22.–24. mája 2013 v Salamance (Španielsko). Článok bol uverejnený v zborníku Advances in Intelligent Systems and Computing.

Anglický preklad návrhu algoritmu

Z dôvodu sprístupnenia kompletného návrhu algoritmu Spamia medzinárodnej odbornej verejnosti bol v júli 2013 publikovaný anglický preklad štúdie o návrhu algoritmu s názvom „Proposal of an algorithm for spam filtering“.

Sumarizačná štúdia

Výstupom šiestej, záverečnej etapy projektu bola sumarizačná štúdia „Projekt Spamia“ http://spamia.slovanet.sk/spamia.pdf. V štúdii boli zhrnuté výsledky dosiahnuté v projekte za celú dobu riešenia. Jednotlivé publikované články a čiastkové štúdie tvoria prílohy záverečnej štúdie.

Popis výstupov za rok 2013:

V roku 2013 boli publikované nasledovné vedecké a odborné články a štúdie:

- „Strong laws for recurrence quantification analysis“, International Journal of Bifurcation and Chaos, vol. 23, no. 8 (2013)
- „OPTICS-based clustering of emails represented by quantitative profiles“, Distributed Computing and Artificial Intelligence, Advances in Intelligent Systems and Computing, Springer (2013)
- „Komparatívny výskum účinnosti algoritmu Spamia“ (2013)
- „Proposal of an algorithm for spam filtering“ (2013)
- „Projekt Spamia“ (2013)