Analyzing Features of Japanese Splogs and Characteristics of Keywords

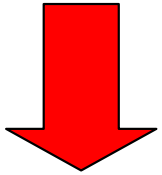
Yuuki Sato[1] **Takehito Utsuro**[1]
Tomohiro Fukuhara[2]
Yasuhide Kawada[3] **Yoshiaki Murakami**[3]
Hiroshi Nakagawa[4] **Noriko Kando**[5]

1 University of Tsukuba, 2,4 University of Tokyo,
3 Navix Co., Ltd., 5 National Institute of Informatics

AIRWeb2008, April 22nd, 2008 @Beijing, CHINA. WWW2008

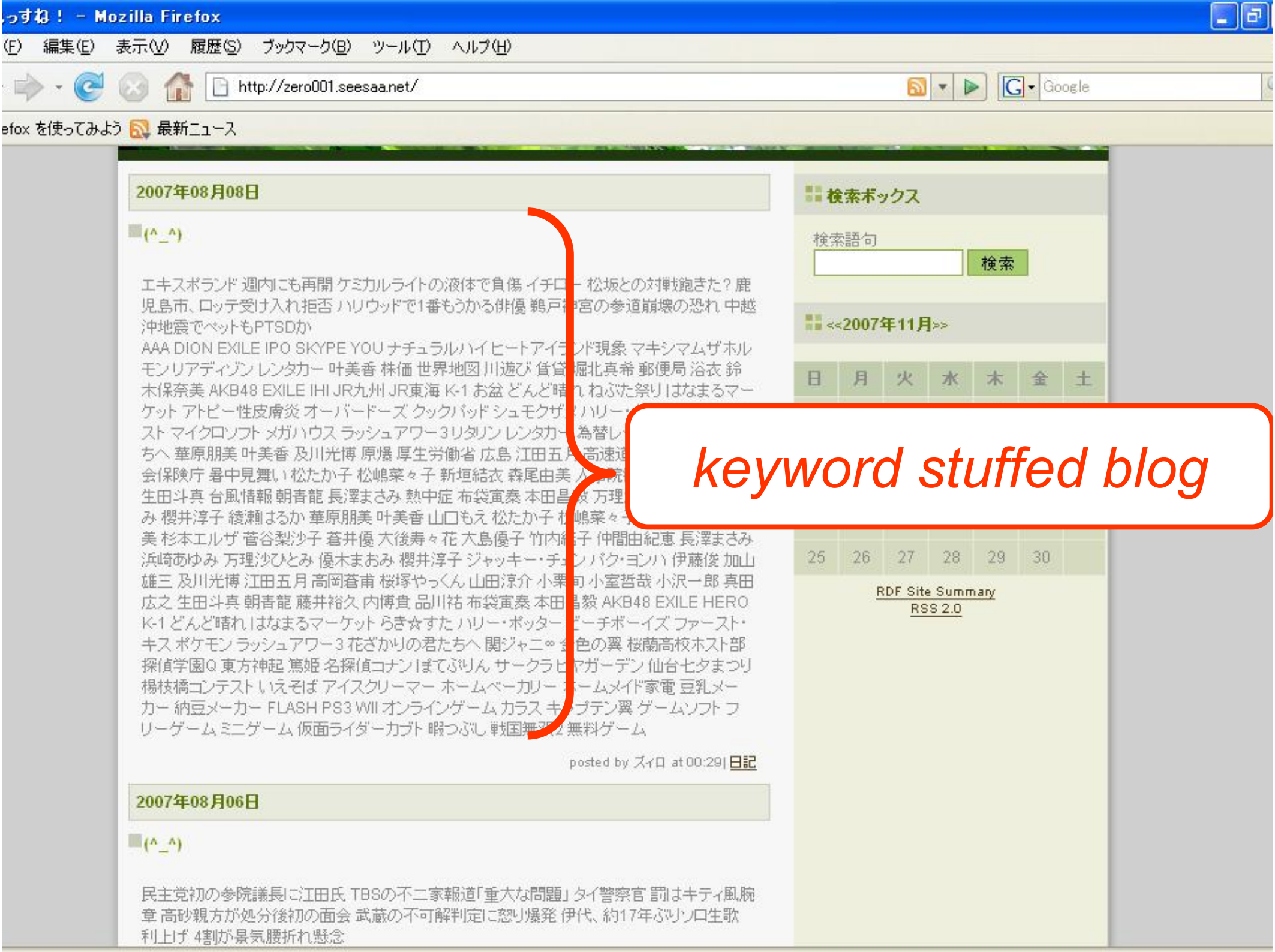
Background (1/2)

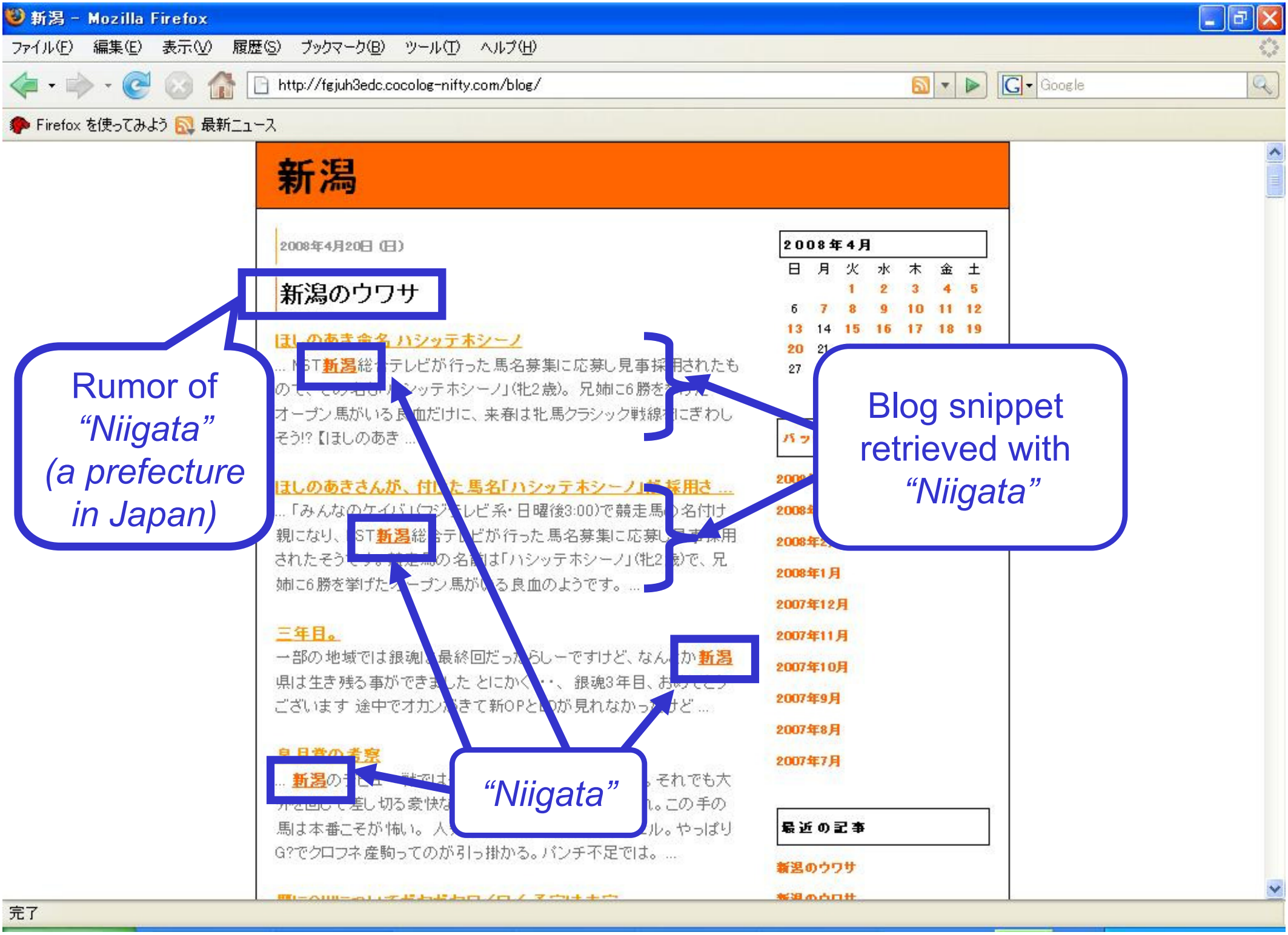
- Opinion Mining from Blogs

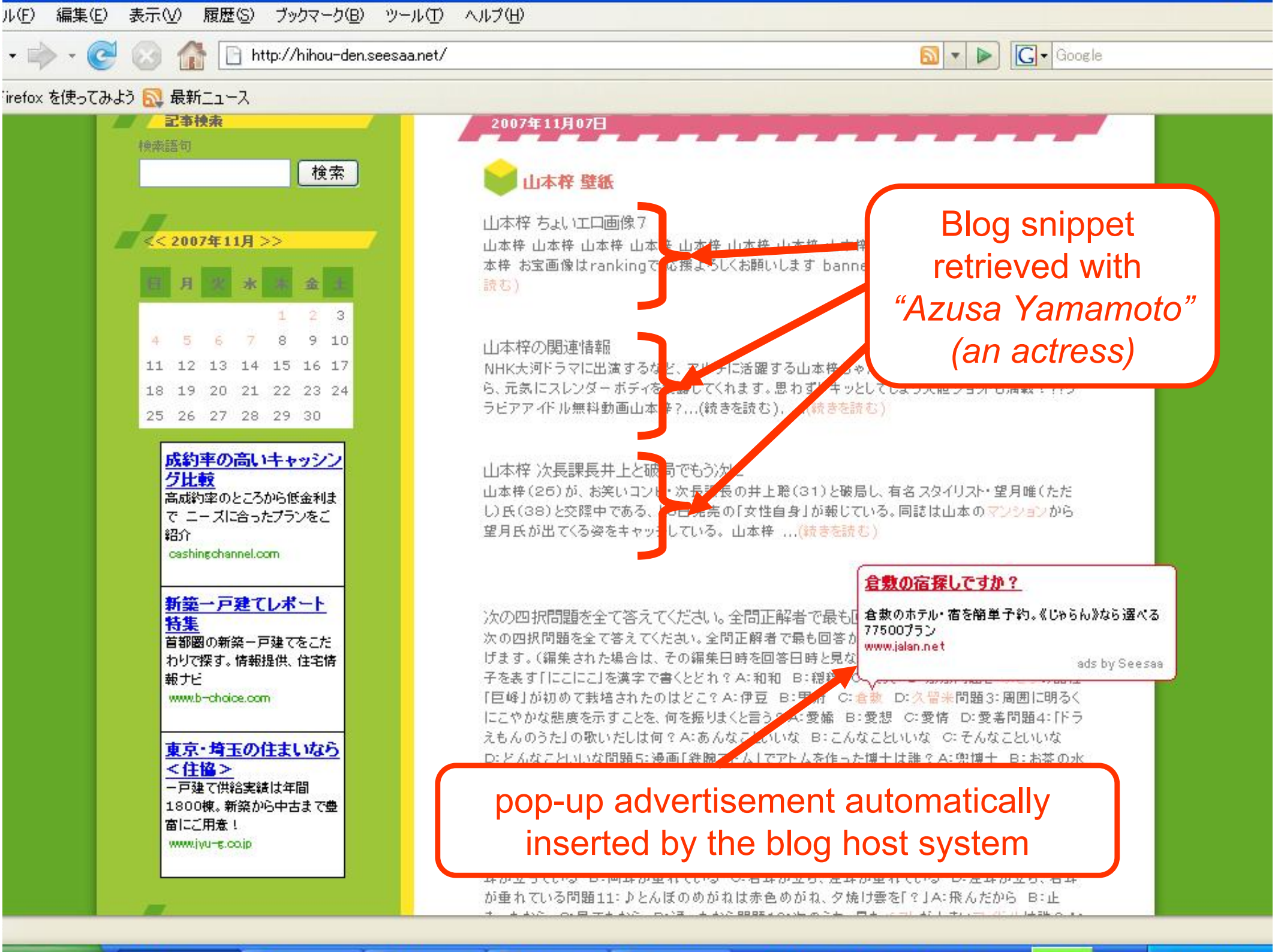


- Splogs are Serious Noise in Opinion Mining
  - e.g., larger scale statistics (2008 Mar.)
    - 40% of Japanese Blog Articles in BuzzPulse, nifty are Splogs, 2007 Oct. ～ 2008 Feb.

- Automatic Detection is highly Expected.

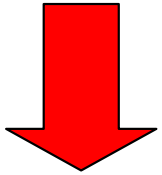

っすね！ - Mozilla Firefox
(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://zero001.seesaa.net/
Google
efox を使ってみよう 最新ニュース
2007年08月08日
(^_^)
エキスポランド 週内にも再開 ケミカルライトの液体で負傷 イチロー 松坂との対戦飽きた？ 鹿児島市、ロッテ受け入れ拒否 ハリウッドで1番もうかる俳優 鵜戸神宮の参道崩壊の恐れ 中越沖地震でペットもPTSDか
AAA DION EXILE IPO SKYPE YOU ナチュラルハイヒートアイランド現象 マキシマムザホルモンリアディゾン レンタカー 叶美香 株価 世界地図 川遊び 賃貸 堀北真希 郵便局 浴衣 鈴木保奈美 AKB48 EXILE IHI JR九州 JR東海 K-1 お盆 どんど晴れ ねぶた祭り はなまるマーケット アトピー性皮膚炎 オーバードーズ クックパッド シュモクザメ ハリー
スト マイクロソフト メガハウス ラッシュアワー3 リタリン レンタカー 為替レ
ちへ 華原朋美 叶美香 及川光博 原爆 厚生労働省 広島 江田五月 高速道
会保険庁 暑中見舞い 松たか子 松嶋菜々子 新垣結衣 森尾由美 人
生田斗真 台風情報 朝青龍 長澤まさみ 熱中症 布袋寅泰 本田昌
み 櫻井淳子 綾瀬はるか 華原朋美 叶美香 山口もえ 松たか子 松
美 杉本エルザ 菅谷梨沙子 蒼井優 大後寿々花 大島優子 竹内結子 仲間由紀恵 長澤まさみ
浜崎あゆみ 万理沙ひとみ 優木まおみ 櫻井淳子 ジャッキー・チェン パク・ヨンハ 伊藤俊 加山雄三 及川光博 江田五月 高岡蒼甫 桜塚やっくん 山田涼介 小栗旬 小室哲哉 小沢一郎 真田広之 生田斗真 朝青龍 藤井裕久 内博貴 品川祐 布袋寅泰 本田昌毅 AKB48 EXILE HERO
K-1 どんど晴れ はなまるマーケット らき☆すた ハリー・ポッター ビーチボーイズ ファースト・キス ポケモン ラッシュアワー3 花ざかりの君たちへ 関ジャニ∞ 金色の翼 桜蘭高校ホスト部 探偵学園Q 東方神起 篤姫 名探偵コナン ぽてぷりん サークラヒアガーデン 仙台七夕まつり 楊枝橋コンテスト いえそば アイスクリーマー ホームベーカリー ホームメイド家電 豆乳メーカー 納豆メーカー FLASH PS3 Wii オンラインゲーム カラス キャプテン翼 ゲームソフト フリーゲーム ミニゲーム 仮面ライダーカブト 暇つぶし 戦国無双2 無料ゲーム
posted by ズイロ at 00:29| 日記
2007年08月06日
(^_^)
民主党初の参院議長に江田氏 TBSの不二家報道「重大な問題」 タイ警察官 罰はキティ風腕章 高砂親方が処分後初の面会 武蔵の不可解判定に怒り爆発 伊代、約17年ぶりソロ生歌 利上げ 4割が景気腰折れ懸念
検索ボックス
検索語句
検索
<<2007年11月>>
日 月 火 水 木 金 土
25 26 27 28 29 30
RDF Site Summary
RSS 2.0
keyword stuffed blog


新潟 - Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://fgjuh3edc.cocolog-nifty.com/blog/
Google
Firefox を使ってみよう
最新ニュース
新潟
2008年4月20日 (日)
新潟のウワサ
2008年4月
日 月 火 水 木 金 土
最近の記事
新潟のウワサ
2008年1月
2007年12月
2007年11月
2007年10月
2007年9月
2007年8月
2007年7月
Rumor of "Niigata" (a prefecture in Japan)
Blog snippet retrieved with "Niigata"
"Niigata"
完了


ル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://hihou-den.seesaa.net/
Google
irefox を使ってみよう 最新ニュース
記事検索
検索語句
検索
<< 2007年11月 >>
日 月 火 水 木 金 土
1 2 3
4 5 6 7 8 9 10
11 12 13 14 15 16 17
18 19 20 21 22 23 24
25 26 27 28 29 30
成約率の高いキャッシング比較
高成約率のところから低金利まで ニーズに合ったプランをご紹介
cashingchannel.com
新築一戸建てレポート特集
首都圏の新築一戸建てをこだわりで探す。情報提供、住宅情報ナビ
www.b-choice.com
東京・埼玉の住まいなら<住協>
一戸建て供給実績は年間1800棟。新築から中古まで豊富にご用意！
www.jyu-s.co.jp
2007年11月07日
山本梓 壁紙
山本梓 ちょいエロ画像7
山本梓 山本梓 山本梓 山本梓
本梓 お宝画像はrankingで応援よろしくお願いします banne
読む)
山本梓の関連情報
NHK大河ドラマに出演するなど、
ら、元気にスレンダーボディを
ラビアアイドル無料動画山本
山本梓 次長課長井上と破局でもう次に
山本梓(26)が、お笑いコンビ・次長課長の井上聡(31)と破局し、有名スタイリスト・望月唯(ただし)氏(38)と交際中である、
望月氏が出てくる姿をキャッチしている。山本梓 ...(続きを読む)
倉敷の宿探しですか？
倉敷のホテル・宿を簡単予約。《じゃらん》なら選べる77500プラン
www.jalan.net
ads by Seesaa
次の四択問題を全て答えてください。全問正解者で最も
次の四択問題を全て答えてください。全問正解者で最も回答
げます。(編集された場合は、その編集日時を回答日時と見な
子を表す「にこにこ」を漢字で書くとどれ？A:和和 B:
「巨峰」が初めて栽培されたのはどこ？A:伊豆 B:
にこやかな態度を示すことを、何を振りまくと言う
えもんのうた」の歌いだしは何？A:あんなこといいな B:こんなことといいな C:そんなことといいな
D:どんなことといいな問題5:漫画「鉄腕
が垂れている問題11:♪とんぼのめがねは赤色めがね、夕焼け雲を「？」A:飛んだから B:止
Blog snippet retrieved with "Azusa Yamamoto" (an actress)
pop-up advertisement automatically inserted by the blog host system

Background (1/2)

- Opinion Mining from Blogs



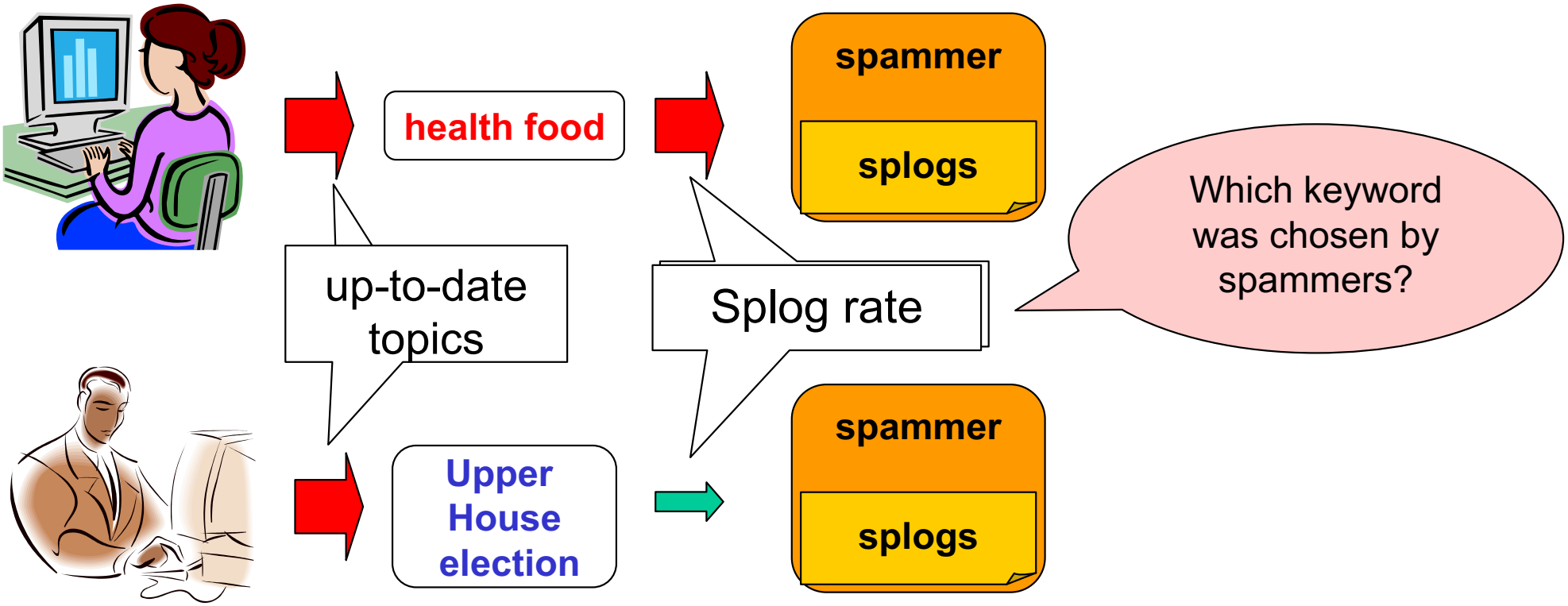
- Splogs are Serious Noise in Opinion Mining
  - e.g., larger scale statistics (2008 Mar.)
    - 40% of Japanese Blog Articles in BuzzPulse, nifty are Splogs, 2007 Oct. ～ 2008 Feb.

- Automatic Detection is highly Expected.

Background (2/2)

- for SEO, Spammers use certain Keywords

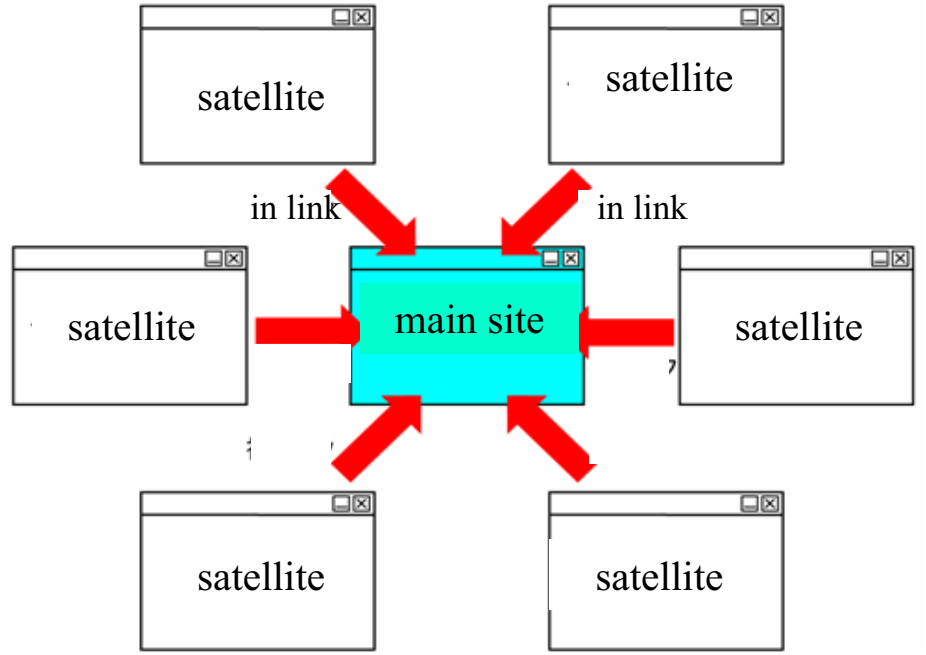
  when Creating Splogs



$50 Software Package for Massive Splog Creation

Featuring

- *SEO*
- *Affiliate Program*



Purpose of this Research

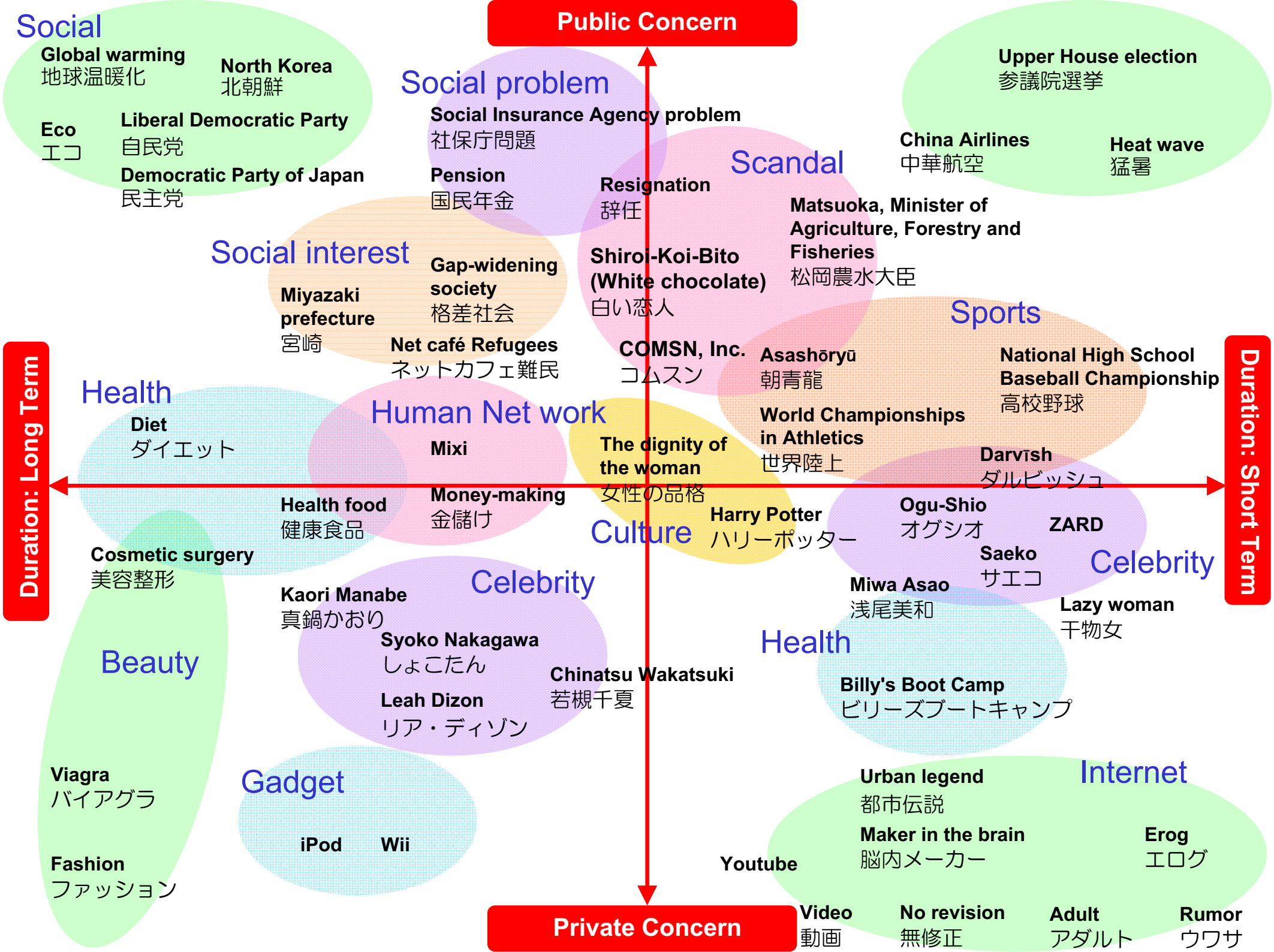
- Manually Analyzing Correlation of Splogs / Splog rates and Keywords included in Japanese Splogs

-

Purpose of this Research

- Manually Analyzing Correlation of Splogs / Splog rates and Keywords included in Japanese Splogs

- Features of Keywords
  - Representing a Topic by a Keyword
    - Topics of Public/Private Concerns
    - Duration Time of a Topic ･･･ with / without Burst
  - Splog rate of Blog Sites including the Keyword
- Features of Splogs
  - Affiliate / Content Source / Creation Procedure
  - Classifying Spammers into *Professional / Amateur*


Public Concern
Private Concern
Duration: Long Term
Duration: Short Term
Social
Global warming
地球温暖化
North Korea
北朝鮮
Eco
エコ
Liberal Democratic Party
自民党
Democratic Party of Japan
民主党
Social problem
Social Insurance Agency problem
社保庁問題
Pension
国民年金
Scandal
Resignation
辞任
Shiroi-Koi-Bito
(White chocolate)
白い恋人
Matsuoka, Minister of Agriculture, Forestry and Fisheries
松岡農水大臣
Upper House election
参議院選挙
China Airlines
中華航空
Heat wave
猛暑
Social interest
Miyazaki prefecture
宮崎
Gap-widening society
格差社会
Net café Refugees
ネットカフェ難民
COMSN, Inc.
コムスン
Sports
Asashōryū
朝青龍
National High School Baseball Championship
高校野球
World Championships in Athletics
世界陸上
Health
Diet
ダイエット
Human Net work
Mixi
Money-making
金儲け
Health food
健康食品
Cosmetic surgery
美容整形
The dignity of the woman
女性の品格
Culture
Harry Potter
ハリーポッター
Darvīsh
ダルビッシュ
Ogu-Shio
オグシオ
ZARD
Saeko
サエコ
Celebrity
Miwa Asao
浅尾美和
Lazy woman
干物女
Celebrity
Kaori Manabe
真鍋かおり
Syoko Nakagawa
しょこたん
Leah Dizon
リア・ディゾン
Chinatsu Wakatsuki
若槻千夏
Health
Billy's Boot Camp
ビリーズブートキャンプ
Beauty
Viagra
バイアグラ
Fashion
ファッション
Gadget
iPod
Wii
Internet
Urban legend
都市伝説
Maker in the brain
脳内メーカー
Erog
エログ
Youtube
Video
動画
No revision
無修正
Adult
アダルト
Rumor
ウワサ

Purpose of this Research

- Manually Analyzing Correlation of Splogs / Splog rates and Keywords included in Japanese Splogs

- Features of Keywords
  - Representing a Topic by a Keyword
    - Topics of Public/Private Concerns
    - Duration Time of a Topic ・・・ with / without Burst
  - Splog rate of Blog Sites including the Keyword
- Features of Splogs
  - Affiliate / Content Source / Creation Procedure
  - Classifying Spammers into *Professional / Amateur*

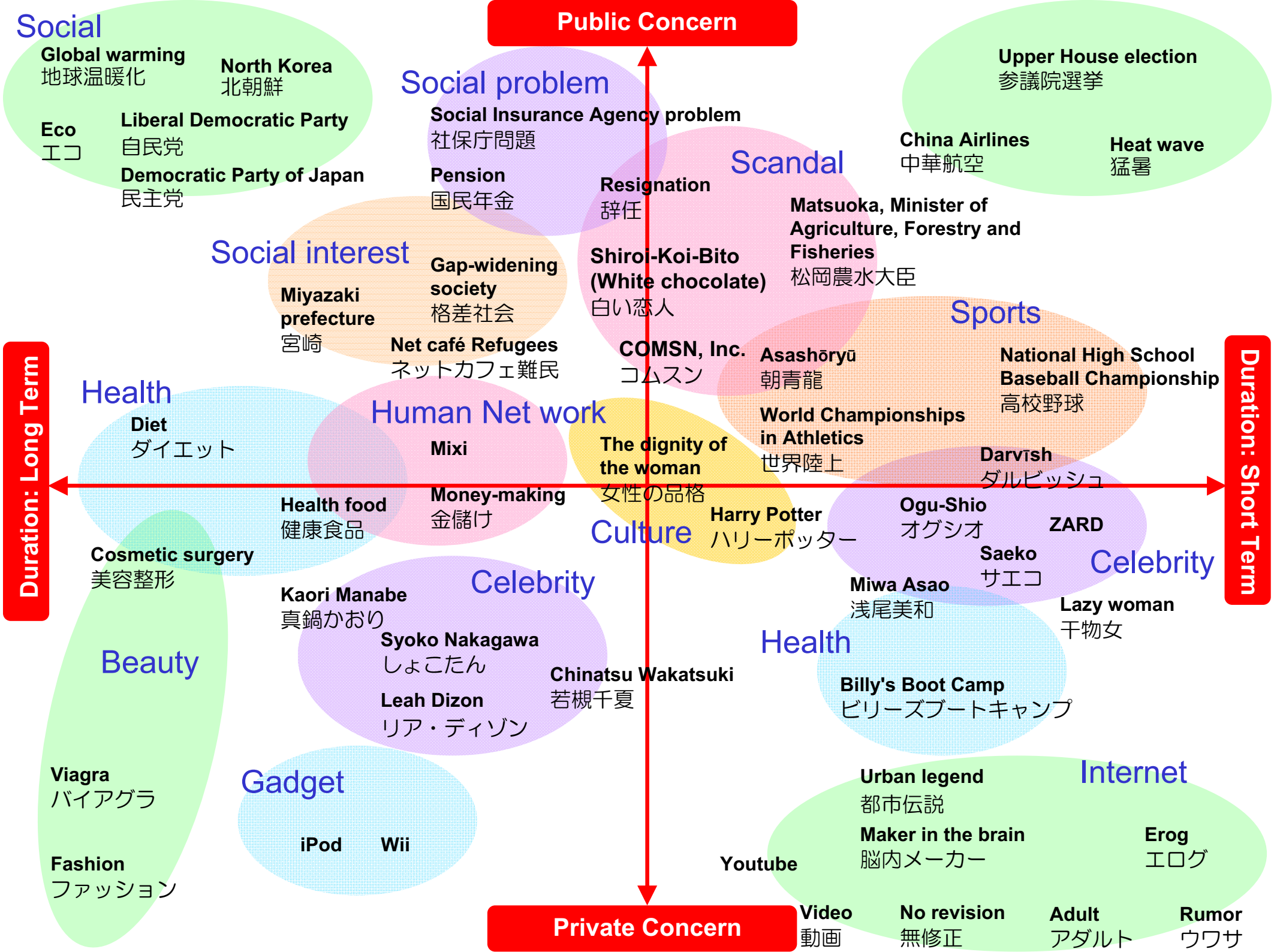
Procedure of Collecting and Annotating Splogs

1. Selecting 50 sample keywords
   balanced on the map.

2. For each keyword, collecting blog site URLs
   including the keyword on its burst date.

3. Sampling blog site URLs
   including those with the most frequent posts.

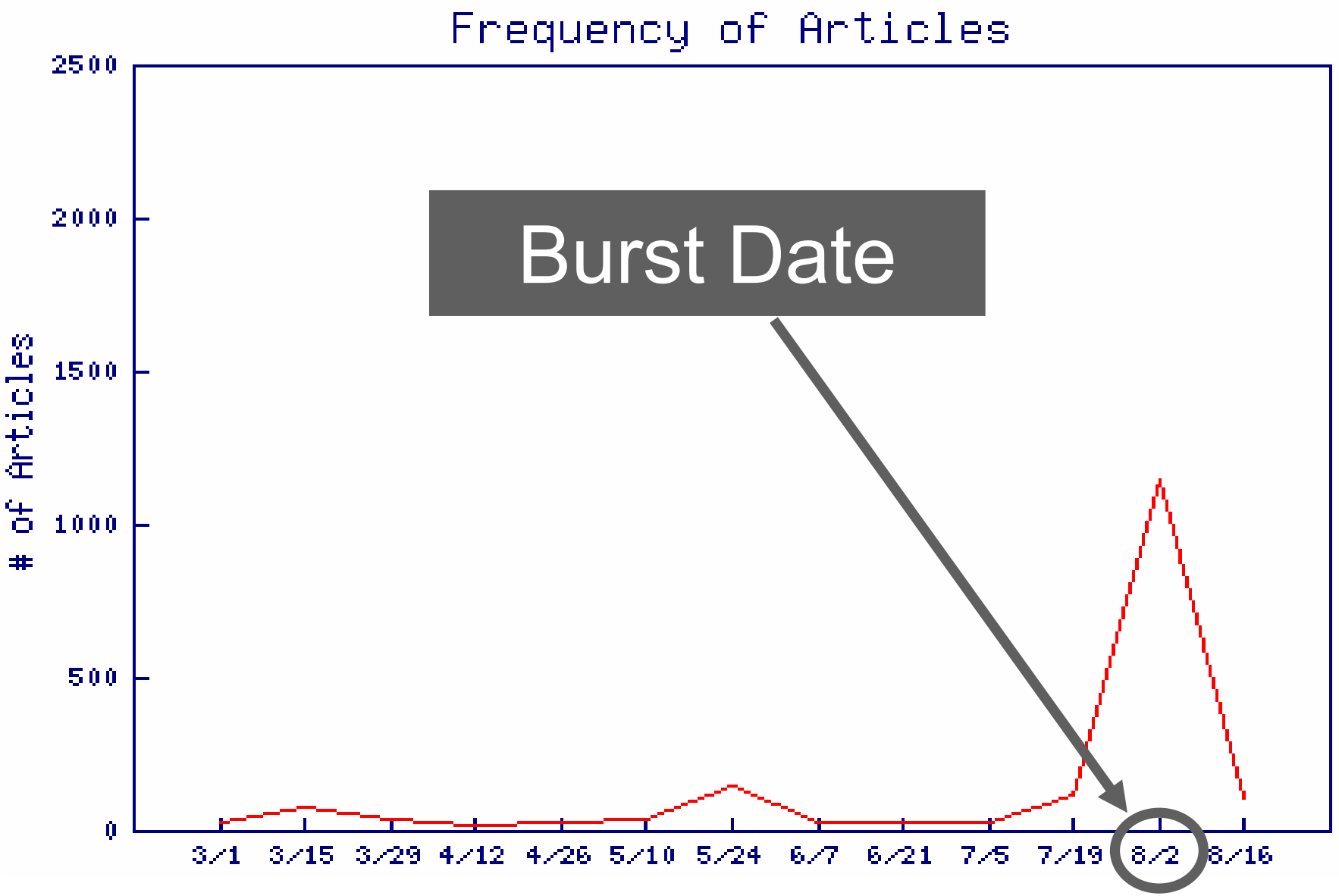
4. Manual assignment of splog features and
   classifying splog/authentic blog.


Public Concern
Private Concern
Duration: Long Term
Duration: Short Term
Social
Global warming
地球温暖化
North Korea
北朝鮮
Eco
エコ
Liberal Democratic Party
自民党
Democratic Party of Japan
民主党
Social problem
Social Insurance Agency problem
社保庁問題
Pension
国民年金
Upper House election
参議院選挙
China Airlines
中華航空
Heat wave
猛暑
Scandal
Resignation
辞任
Matsuoka, Minister of Agriculture, Forestry and Fisheries
松岡農水大臣
Shiroi-Koi-Bito (White chocolate)
白い恋人
Social interest
Gap-widening society
格差社会
Miyazaki prefecture
宮崎
Net café Refugees
ネットカフェ難民
COMSN, Inc.
コムスン
Sports
Asashōryū
朝青龍
National High School Baseball Championship
高校野球
World Championships in Athletics
世界陸上
Darvīsh
ダルビッシュ
Health
Diet
ダイエット
Human Net work
Mixi
The dignity of the woman
女性の品格
Health food
健康食品
Money-making
金儲け
Harry Potter
ハリーポッター
Culture
Ogu-Shio
オグシオ
ZARD
Saeko
サエコ
Celebrity
Cosmetic surgery
美容整形
Celebrity
Kaori Manabe
真鍋かおり
Miwa Asao
浅尾美和
Lazy woman
干物女
Syoko Nakagawa
しょこたん
Chinatsu Wakatsuki
若槻千夏
Health
Billy's Boot Camp
ビリーズブートキャンプ
Beauty
Leah Dizon
リア・ディゾン
Viagra
バイアグラ
Gadget
Internet
Urban legend
都市伝説
Maker in the brain
脳内メーカー
Erog
エログ
iPod
Wii
Youtube
Fashion
ファッション
Video
動画
No revision
無修正
Adult
アダルト
Rumor
ウワサ

Procedure of Collecting and Annotating Splogs

1. Selecting 50 sample keywords balanced on the map.
2. For each keyword, collecting blog site URLs including the keyword on its burst date.
3. Sampling blog site URLs including those with the most frequent posts.
4. Manual assignment of splog features and classifying splog/authentic blog.

Collecting blog site URLs including the keyword on its burst date



Procedure of Collecting and Annotating Splogs

1. Selecting 50 sample keywords
   balanced on the map.

2. For each keyword, collecting blog site URLs
   including the keyword on its burst date.

3. Sampling blog site URLs
   including those with the most frequent posts.

4. Manual assignment of splog features and
   classifying splog/authentic blog.

Features for Characterizing Splogs and Rate in Splogs

Affiliate Features	A1: links to affiliated sites	80.5%
	A2: advertisement articles (posts)	31.0%
	A3: articles (posts) with adult content	8.1%
	A4: keywords with popup advertisement	42.1%
Content Source Features	S1: excerpt from news articles	14.3%
	S2: excerpt from blog articles (posts) or other web texts	70.8%
	S3: excerpt from advertisement pages	27.1%
	S4: originally written texts	2.9%
	S5: meaningless sequence of words	3.6%
Creation Procedure Features	P1: excerpt from other sources, selected without keyword retrieval	12.7%
	P2: excerpt from other sources, retrieved with a keyword varying day by day	49.5%
	P3: excerpt from other sources, retrieved with a single keyword throughout a blog homepage	36.9%
	P4: keyword stuffed blog	11.5%
	P5: automatically generated text	4.5%

Manual Analysis of Splogs

- Blog Host Distribution
- Classifying Spammers into *Professional / Amateur*
  - Analyzing Professional Spammers
- Splog Features and Keywords
- Correlation:
  - Characteristics of Keywords:
    - Public / Private Concern
  - Splog Rate per Keyword
    - Professional Spammer Rate
    - Amateur Only Splog Rate

Manual Analysis of Splogs

- Blog Host Distribution 
- Classifying Spammers into *Professional / Amateur*
  - Analyzing Professional Spammers
- Splog Features and Keywords
- Correlation:
  - Characteristics of Keywords:
    - Public / Private Concern
  - Splog Rate per Keyword
    - Professional Spammer Rate
    - Amateur Only Splog Rate

Splog Host Statistics (1/2)

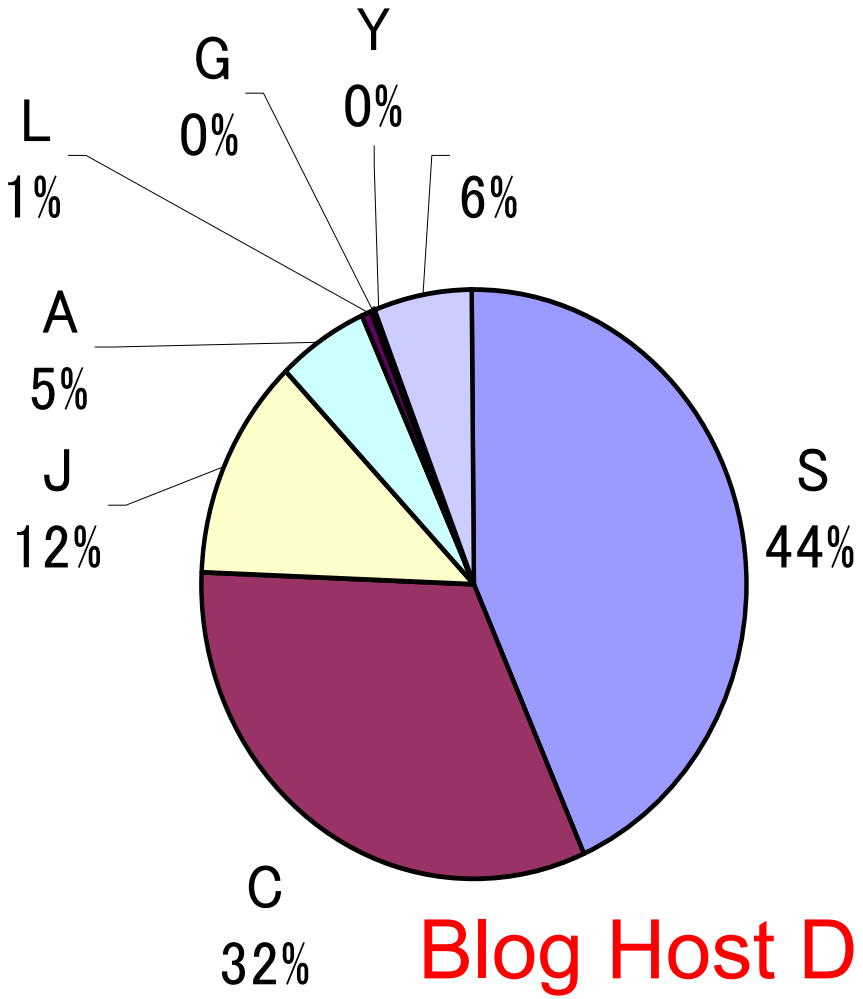
- for 22 / 50 keywords, 2145 blog sites

- for the hosts S and C, splog rates
  in our blog data set are around 50%,
  paying less costs of manually removing splogs.

Splog Rate per Blog Host

Host	S	C	J	A	L	G	Y	rest	**total**
splog	192	142	54	24	3	1	0	26	442
non splog	203	115	169	355	128	130	207	296	1703
splog rate(%)	48.6	55.3	24.2	6.3	2.3	0.8	0.0	8.7	20.6

Splog Host Statistics (2/2)



Blog Host Distribution in Splogs

- for 22 / 50 keywords, 2145 blog sites

- 88% of splogs from top-3 hosts (S,C,J)

Manual Analysis of Splogs

- Blog Host Distribution
- Classifying Spammers into *Professional / Amateur*
  - Analyzing Professional Spammers 
- Splog Features and Keywords
- Correlation:
  - Characteristics of Keywords:
    - Public / Private Concern
  - Splog Rate per Keyword
    - Professional Spammer Rate
    - Amateur Only Splog Rate

Professional / Amateur Spammers

Professional Spammer:
who created more than splog sites in our data set.

Amateur Spammer:
who created only one splog site in our data set.

- Professional Spammers are manually identified by comparing the similarity of HTML structures.

- 60% of Splogs are created by 10 Professional Spammers

10 Professional Spammers in our Splog Data Set

ID	# of Splogs	Affiliate Features	Content Source Features	Creation Procedure Features	keywords
1	115 (42.3%)	A1: links to affiliated sites, A4: popup advertisement	S2: blog or other web texts	P3: retrieved with a single keyword	**rumor**, no revision, cosmetic surgery, **Asasho-ryu**, Saeko, China Airlines, COMSN, Inc., ZARD, heat wave, Wii, North Korea, "lazy woman"
2	56 (20.6%)	A1: links to affiliated sites	S2: blog or other web texts	P2: retrieved with a keyword varying day by day	**Erog**
3	30 (11.0%)	A1: links to affiliated sites	S1: news articles, S3: advertisement pages	P1: selected without keyword retrieval	**national pension**, COMSN, Inc.
4	26 (9.6%)	A1: links to affiliated sites, A2: advertisement articles, A4: popup advertisement	S2: blog or other web texts, S3: advertisement pages	P2: retrieved with a keyword varying day by day	**national pension**
5	20 (7.4%)	A1: links to affiliated sites, A2: advertisement articles	S3: advertisement pages	P2: retrieved with a keyword varying day by day, P4: keyword stuffed blog	**health food**
6	10 (3.7%)	A1: links to affiliated sites, A3: adult content, A4: popup advertisement	S1: news articles, S2: blog or other web texts	P1: selected without keyword retrieval	**Erog, Asasho-ryu,**
7‥‧10	15 (5.5%)	---	---	---	**Erog, health food**, Viagra, cosmetic surgery,
Total	272 (100%)	---	---	---	---

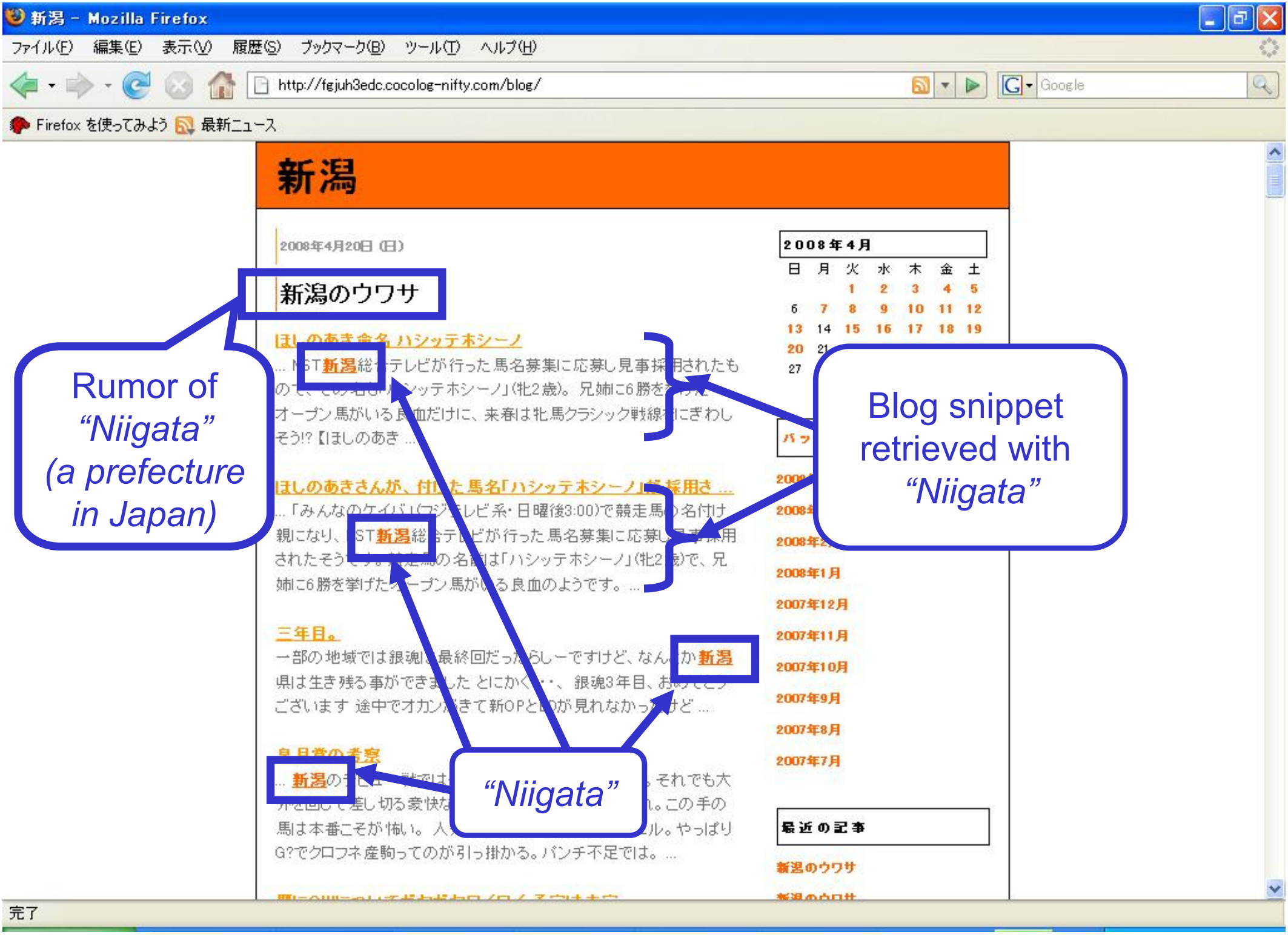
Manual Analysis of Splogs

- Blog Host Distribution
- Classifying Spammers into *Professional / Amateur*
  - Analyzing Professional Spammers
- Splog Features and Keywords 
- Correlation:
  - Characteristics of Keywords:
    - Public / Private Concern
  - Splog Rate per Keyword
    - Professional Spammer Rate
    - Amateur Only Splog Rate

10 Professional Spammers in our Splog Data Set

ID	# of Splogs	Affiliate Features	Content Source Features	Creation Procedure Features	keywords
1	115 (42.3%)	A1: links to affiliated sites, A4: popup advertisement	S2: blog or other web texts	P3: retrieved with a single keyword	**rumor**, no revision, cosmetic surgery, **Asasho-ryu**, Saeko, China Airlines, COMSN, Inc., ZARD, heat wave, Wii, North Korea, "lazy woman"
2	56 (20.6%)	A1: links to affiliated sites	S2: [illegible]	P2: retrieved with a	
3	30 (11.0%)	A1: links to affiliated sites	S3: advertisement pages	[illegible] keyword retrieval	**national pension**, COMSN, Inc.
4	26 (9.6%)	A1: links to affiliated sites, A2: advertisement articles, A4: popup advertisement	S2: blog or other web texts, S3: advertisement pages	P2: retrieved with a keyword varying day by day	**national pension**
5	20 (7.4%)	A1: links to affiliated sites, A2: advertisement articles	S3: advertisement pages	P2: retrieved with a keyword varying day by day, P4: keyword stuffed blog	**health food**
6	10 (3.7%)	A1: links to affiliated sites, A3: adult content, A4: popup advertisement	S1: news articles, S2: blog or other web texts	P1: selected without keyword retrieval	**Erog, Asasho-ryu,**
7・・・10	15 (5.5%)	---	---	---	**Erog, health food**, Viagra, cosmetic surgery,
Total	272 (100%)	---	---	---	---

Splog of "rumor of *X*" type


新潟 - Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://fgjuh3edc.cocolog-nifty.com/blog/
Google
Firefox を使ってみよう
最新ニュース
新潟
2008年4月20日 (日)
新潟のウワサ
2008年4月
日 月 火 水 木 金 土
最近の記事
新潟のウワサ
2008年1月
2007年12月
2007年11月
2007年10月
2007年9月
2007年8月
2007年7月
Rumor of "Niigata" (a prefecture in Japan)
Blog snippet retrieved with "Niigata"
"Niigata"
完了

10 Professional Spammers in our Splog Data Set

ID	# of Splogs	Affiliate Features	Content Source Features	Creation Procedure Features	keywords
1	115 (42.3%)	A1: links to affiliated sites, A4: popup advertisement	S2: blog or other web texts	P3: retrieved with a single keyword	**rumor**, no revision, cosmetic surgery, **Asasho-ryu**, Saeko, China Airlines, COMSN, Inc., ZARD, heat wave, Wii, North Korea, "lazy woman"
2	56 (20.6%)	A1: links to affiliated sites	S2: blog or other web texts	P2: retrieved with a keyword varying day by day	**Erog**
3	30 (11.0%)	A1: links to affiliated sites	S1: news articles, S3: advertisement pages	P1: selected without keyword retrieval	**national pension**, COMSN, Inc.
4		A1: links to affiliated	S2: blog or	P2: retrieved with a	
5		articles	pages	stuffed blog	
6	10 (3.7%)	A1: links to affiliated sites, A3: adult content, A4: popup advertisement	S1: news articles, S2: blog or other web texts	P1: selected without keyword retrieval	**Erog, Asasho-ryu,**
7・・・10	15 (5.5%)	---	---	---	**Erog, health food**, Viagra, cosmetic surgery,
Total	272 (100%)	---	---	---	---

- Text content is excerpted from the news articles of the date on which the splog article are created, without keyword retrieval.
- Keywords with *public concern*, "national pension" and "Asasho-ryu" are included in splogs.

Manual Analysis of Splogs

- Blog Host Distribution
- Classifying Spammers into *Professional / Amateur*
  - Analyzing Professional Spammers
- Splog Features and Keywords
- Correlation:

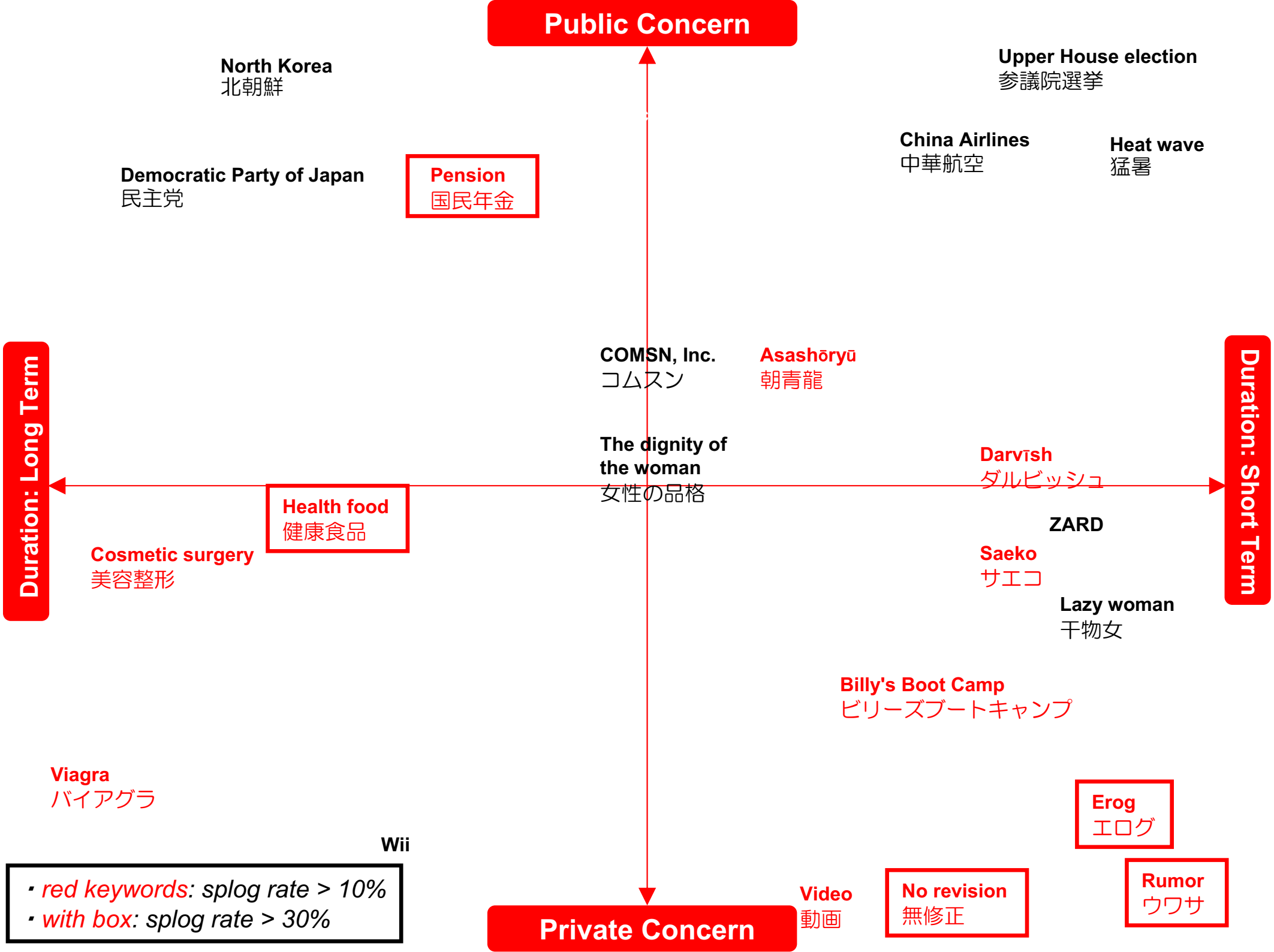
  - Characteristics of Keywords:
    - Public / Private Concern
  - Splog Rate per Keyword
    - Professional Spammer Rate
    - Amateur Only Splog Rate

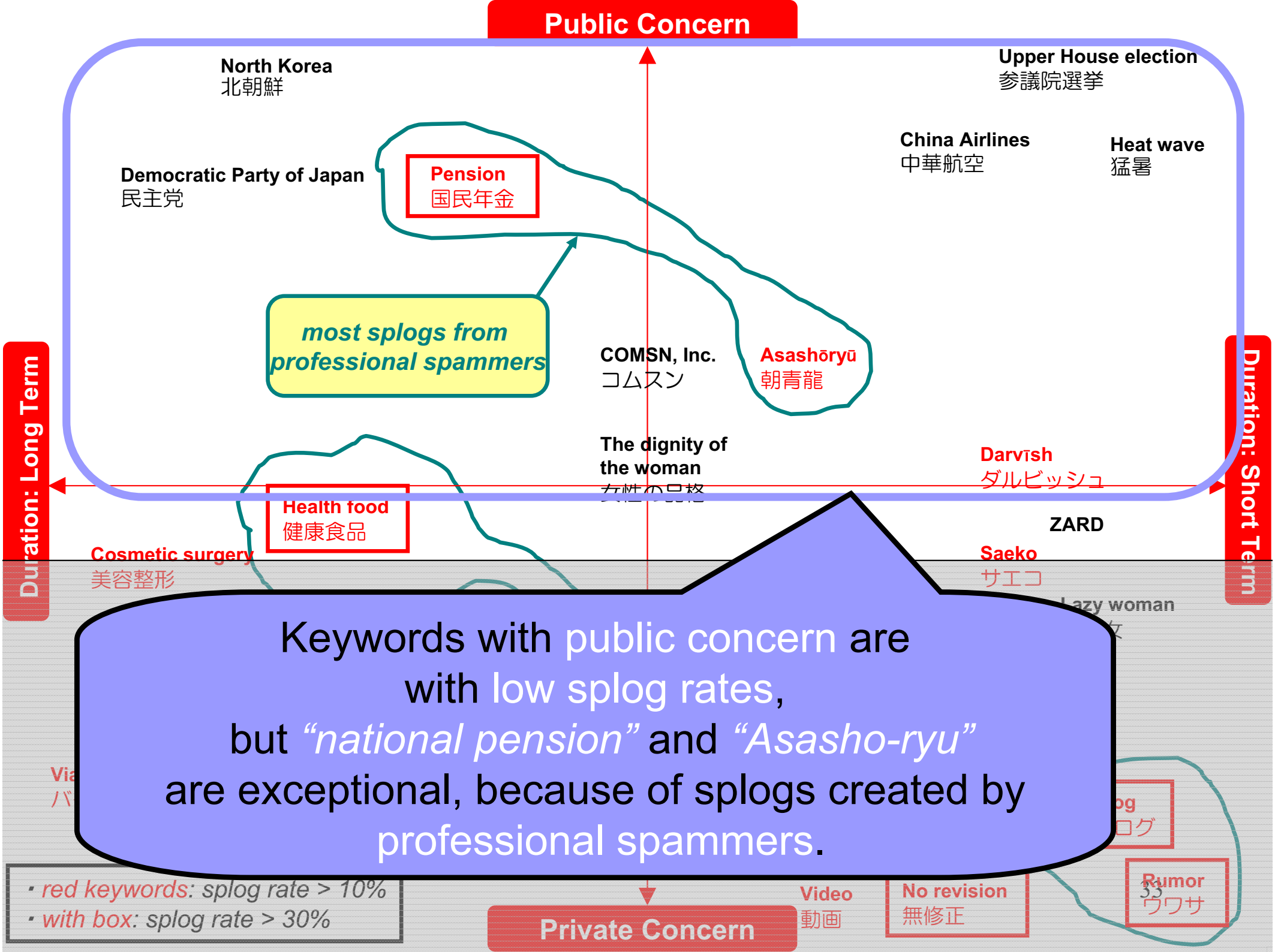
Splog Rate per Keyword: for 22 Keywords

	/ (splog + non-splog))
エログ (Erog, adult)	89.2%
ウワサ (rumor)	88.1%
国民年金 (national pension)	58.1%
無修正 (no revision)	40.9%
健康食品 (health food)	37.4%
美容整形 (comsmetic surgery)	24.4%
バイアグラ (Viagra)	22.5%
ダルビッシュ (Darvish, baseball player)	22.1%
動画 (video)	19.1%
朝青龍 (Asasho-ryu, sumo wrestler)	15.2%
ビリーズブートキャンプ (Billy's Boot Camp)	15.1%
サエコ (Saeko, Darvish's wife)	14.3%
コムスン (COMSN, Inc.)	6.9%
ＺＡＲＤ (singer)	4.7%
中華航空 (China Airlines)	4.7%
北朝鮮 (North Korea)	2.9%
Ｗｉｉ (video game console)	2.8%
猛暑 (heat wave)	2.8%
女性の品格 ("The dignity of the woman", book)	2.0%
干物女 (slang word for "lazy woman")	1.8%
参議院選挙 (Upper House election)	0.0%
民主党 (Democratic Party of Japan)	0.0%
Total	23.5%

90%～30%

30%～10%


Public Concern
North Korea
北朝鮮
Upper House election
参議院選挙
China Airlines
中華航空
Heat wave
猛暑
Democratic Party of Japan
民主党
Pension
国民年金
Duration: Long Term
COMSN, Inc.
コムスン
Asashōryū
朝青龍
The dignity of the woman
女性の品格
Darvīsh
ダルビッシュ
Duration: Short Term
Health food
健康食品
ZARD
Cosmetic surgery
美容整形
Saeko
サエコ
Lazy woman
干物女
Billy's Boot Camp
ビリーズブートキャンプ
Viagra
バイアグラ
Erog
エログ
Wii
・red keywords: splog rate > 10%
・with box: splog rate > 30%
Video
動画
No revision
無修正
Rumor
ウワサ
Private Concern


Public Concern
North Korea
北朝鮮
Upper House election
参議院選挙
China Airlines
中華航空
Heat wave
猛暑
Democratic Party of Japan
民主党
Pension
国民年金
most splogs from professional spammers
COMSN, Inc.
コムスン
Asashōryū
朝青龍
Duration: Long Term
Duration: Short Term
The dignity of the woman
Darvīsh
ダルビッシュ
Health food
健康食品
ZARD
Cosmetic surgery
美容整形
Saeko
サエコ
Lazy woman
Keywords with public concern are with low splog rates, but "national pension" and "Asasho-ryu" are exceptional, because of splogs created by professional spammers.
・red keywords: splog rate > 10%
・with box: splog rate > 30%
Video
動画
No revision
無修正
Rumor
ウワサ
Private Concern
33

Splog Rate per Keyword: for 22 Keywords

	/ (splog + non-splog))
エログ (Erog, adult)	89.2%
ウワサ (rumor)	88.1%
国民年金 (national pension)	58.1%
無修正 (no revision)	40.9%
健康食品 (health food)	37.4%
美容整形 (comsmetic surgery)	24.4%
バイアグラ (Viagra)	22.5%
ダルビッシュ (Darvish, baseball player)	22.1%
動画 (video)	19.1%
朝青龍 (Asasho-ryu, sumo wrestler)	15.2%
ビリーズブートキャンプ (Billy's Boot Camp)	15.1%
サエコ (Saeko, Darvish's wife)	14.3%
[illegible] (COMSN, Inc.)	[illegible]
ＺＡＲＤ (singer)	4.7%
中華航空 (China Airlines)	4.7%
北朝鮮 (North Korea)	2.9%
Ｗｉｉ (video game console)	2.8%
猛暑 (heat wave)	2.8%
女性の品格 ("The dignity of the woman", book)	2.0%
干物女 (slang word for "lazy woman")	1.8%
参議院選挙 (Upper House election)	0.0%
民主党 (Democratic Party of Japan)	0.0%
Total	23.5%

90%～30%

30%～10%

Splog Rate, Professional Spammer Rate, Amateur Only Splog Rate

Kewords	Splog rate (splog / (splog + non-splog))	Professional spammer rate (from professional spammer / splog)	Professional Spammer ID	Amateur Only splog rate (from amateur spammer / (from amateur spammer + non-splog))
エログ (Erog, adult)	89.2%	92.4%	2, 6, 8	38.5%
ウワサ (rumor)	88.1%	94.8%	1	27.8%
国民年金 (national pension)	58.1%	90.2%	3, 4	12.0%
無修正 (no revision)	40.9%	18.5%	1	36.1%
健康食品 (health food)	2	58.7%	5, 7	19.8%
美容整形 (comsmetic surgery)	4%	14.3%	1, 10	21.7%
バイアグラ (Viagra)	5%	11.1%	9	20.5%
ダルビッシュ (Darvish, baseball player)	22.1%	0.0%	---	22.1%
動画 (video)	19.1%	0.0%	---	19.1%
朝青龍 (Asasho-ryu, sumo wrestler)	15.2%	80.0%	1, 6	3.4%
ビリーズブートキャンプ (Billy's Boot Camp)	15.1%	0.0%	---	15.1%
サエコ (Saeko, Darvish's wife)	14.3%	14.3%	1	12.2%
Total of 22 keywords	23.5%	61.5%	10 Groups	9.0%

Splog Rate, Professional Spammer Rate, Amateur Only Splog Rate

Kewords	Splog rate (splog / (splog + non-splog))	Professional spammer rate (from professional spammer / splog)	Professional Spammer ID	Amateur Only splog rate (from amateur spammer / (from amateur spammer + non-splog))
エログ (Erog, adult)	89.2%	92.4%	2, 6, 8	38.5%
ウワサ (rumor)	88.1%	94.8%	1	27.8%
国民年金 (national pension)	58.1%	90.2%	3, 4	12.0%
無修正 (no revision)	40.9%	12.5%	1	36.1%
健康食品 (health food)	37.4%	[illegible]%	5, 7	19.8%
美容整形 (comsmetic surgery)	24.4%	14.3%	1, 10	21.7%
バイアグラ (Viagra)	22.5%	11.1%	9	20.5%
ダルビッシュ (Darvish, baseball player)	22.1%	0.0%	---	22.1%
動画 (video)	19.1%	0.0%	---	19.1%
朝青龍 (Asasho-ryu, sumo wrestler)	15.2%	80.0%	1, 6	3.4%
ビリーズブートキャンプ (Billy's Boot Camp)	15.1%	0.0%	---	15.1%
サエコ (Saeko, Darvish's wife)	14.3%	14.3%	1	12.2%
Total of 22 keywords	23.5%	61.5%	10 Groups	9.0%

Splog Rate, Professional Spammer Rate, Amateur Only Splog Rate

Kewords	Splog rate (splog / (splog + non-splog))	Professional spammer rate (from professional spammer / splog)	Professional Spammer ID	Amateur Only splog rate (from amateur spammer / (from amateur spammer + non-splog))
エログ (Erog, adult)	89.2%	92.4%	2, 6, 8	38.5%
ウワサ (rumor)	88.1%	94.8%	1	27.8%
国民年金 (national pension)	58.1%	90.2%	3, 4	12.0%
無修正 (no revision)	40.9%	18.5%	1	36.1%
健康食品 (health food)	37.4%	58.7%	5, 7	19.8%
美容整形 (comsmetic surgery)	24.4%	14.3%	1, 10	21.7%
バイアグラ (Viagra)	22.5%	11.1%	9	20.5%
ダルビッシュ (Darvish, baseball player)	22.1%	0.0%	---	22.1%
動画 (video)	19.1%	0.0%	---	19.1%
朝青龍 (Asasho-ryu, sumo wrestler)	15.2%	80.0%	1, 6	3.4%
ビリーズブートキャンプ (Billy's Boot Camp)	15.1%	0.0%	---	15.1%
サエコ (Saeko, Darvish's wife)	14.3%	14.3%	1	12.2%
Total of 22 keywords	23.5%	61.5%	10 Groups	9.0%

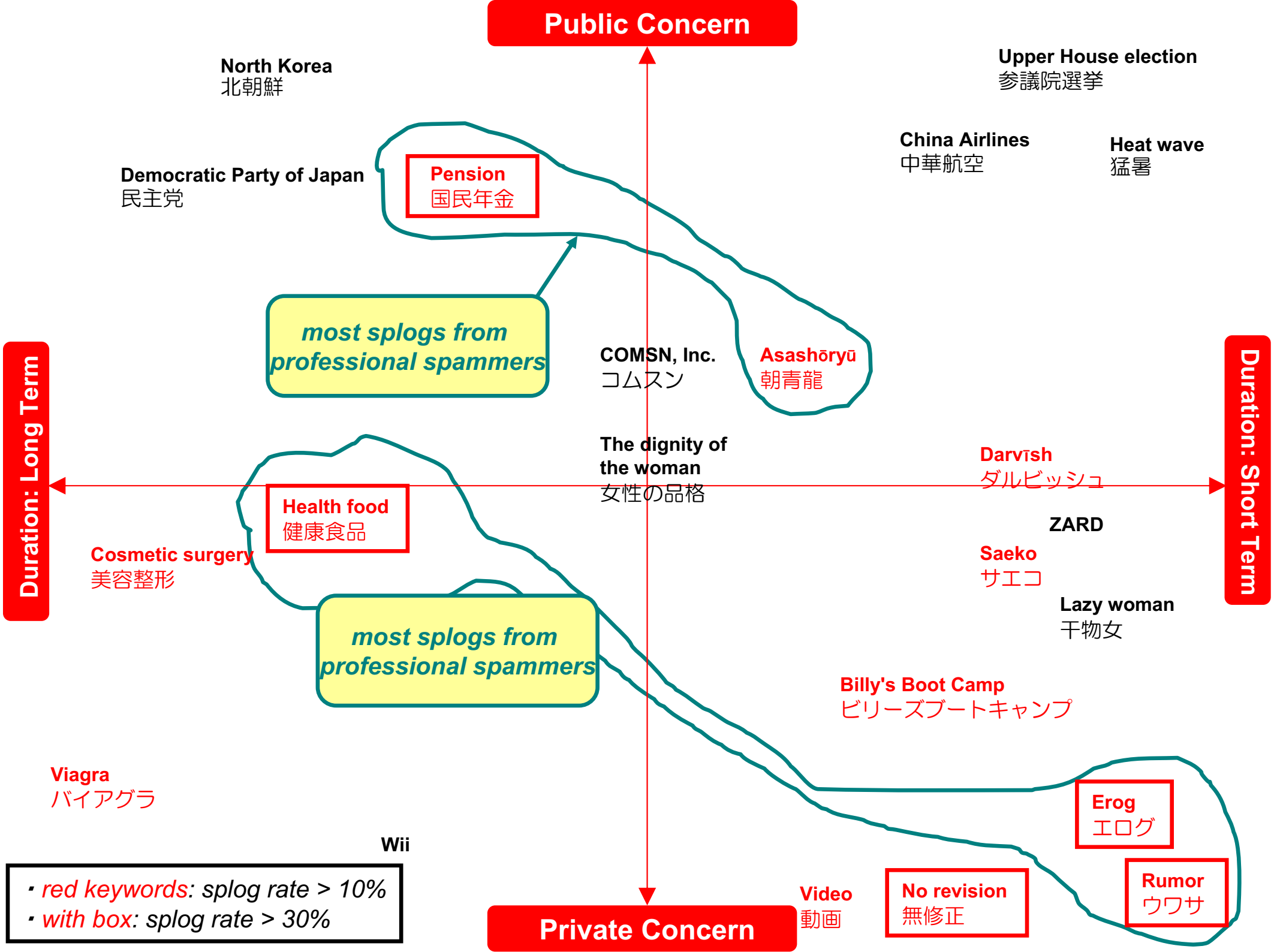
Splog Rate, Professional Spammer Rate, Amateur Only Splog Rate

Kewords	Splog rate (splog / (splog + non-splog))	Professional spammer rate (from professional spammer / splog)	Professional Spammer ID	Amateur Only splog rate (from amateur spammer / (from amateur spammer + non-splog))
エログ (Erog, adult)	89.2%	**92.4%**	**2, 6, 8**	38.5%
ウワサ (rumor)	88.1%	**94.8%**	**1**	27.8%
国民年金 (national pension)	58.1%	**90.2%**	**3, 4**	12.0%
無修正 (no revision)	40.9%	18.5%	**1**	36.1%
健康食品 (health food)	37.4%	**58.7%**	**5, 7**	19.8%
美容整形 (comsmetic surgery)	24.4%	14.3%	**1, 10**	21.7%
バイアグラ (Viagra)	22.5%	11.1%	**9**	20.5%
ダルビッシュ (Darvish, baseball player)	22.1%	0.0%	**---**	22.1%
動画 (video)	19.1%	0.0%	**---**	19.1%
朝青龍 (Asasho-ryu, sumo wrestler)	15.2%	**80.0%**	**1, 6**	3.4%
ビリーズブートキャンプ (Billy's Boot Camp)	15.1%	0.0%	**---**	15.1%
サエコ (Saeko, Darvish's wife)	14.3%	14.3%	**1**	12.2%
Total of 22 keywords	23.5%	**61.5%**	**10 Groups**	**9.0%**

Splog Rate, Professional Spammer Rate, Amateur Only Splog Rate

Kewords	Splog rate (splog / (splog + non-splog))	Professional spammer rate (from professional spammer / splog)	Professional Spammer ID	Amateur Only splog rate (from amateur spammer / (from amateur spammer + non-splog))
エログ (Erog, adult)	89.2%	**92.4%**	**2, 6, 8**	38.5%
ウワサ (rumor)	88.1%	**94.8%**	**1**	27.8%
国民年金 (national pension)	58.1%	**90.2%**	**3, 4**	12.0%
無修正 (no revision)	40.9%	18.5%	**1**	36.1%
健康食品 (health food)	37.4%	**58.7%**	**5, 7**	19.8%
美容整形 (comsmetic surgery)	24.4%	14.3%		21.7%
バイアグラ (Viagra)	22.5%	11.1%		
ダルビッシュ (Darvish, baseball player)	22.1%	0.0%		
動画 (video)	19.1%	0.0%		
朝青龍 (Asasho-ryu, sumo wrestler)	15.2%	**80.0%**		
ビリーズブートキャンプ (Billy's Boot Camp)	15.1%	0.0%	---	15.1%
サエコ (Saeko, Darvish's wife)	14.3%	14.3%	**1**	12.2%
Total of 22 keywords	23.5%	**61.5%**	**10 Groups**	**9.0%**

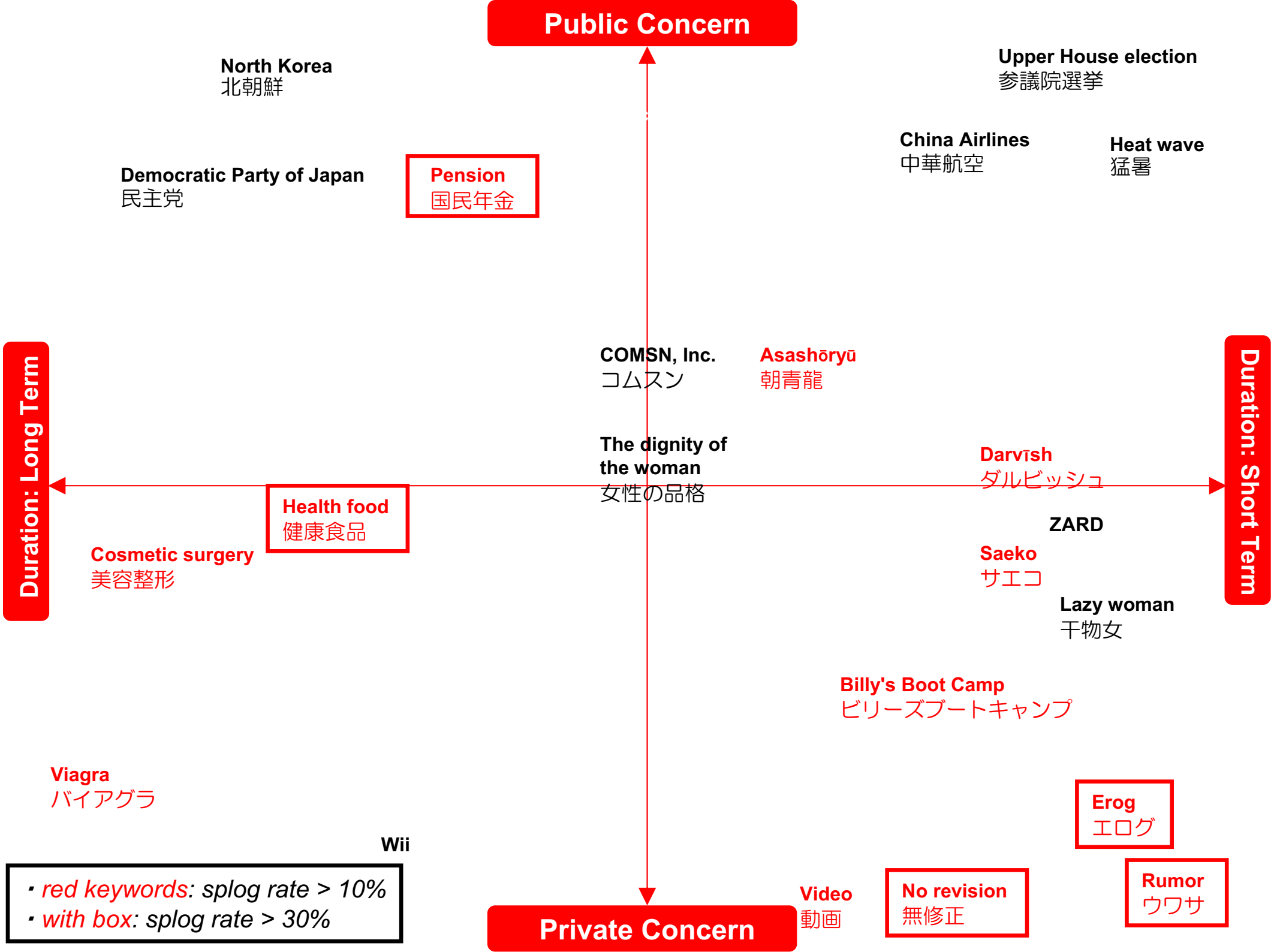
Keywords with high splog rates are related to splogs from professional spammers.

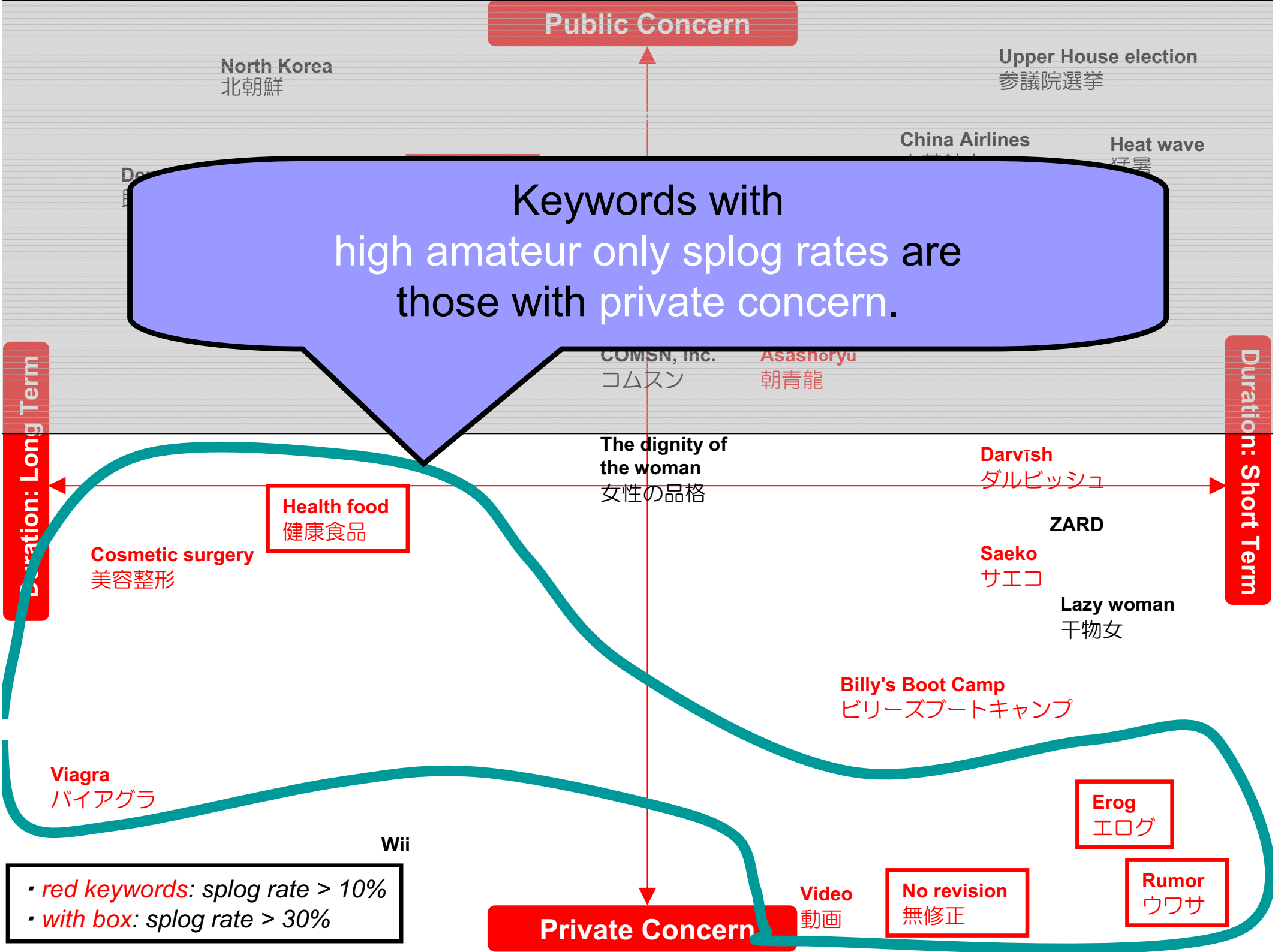

Public Concern
North Korea
北朝鮮
Upper House election
参議院選挙
China Airlines
中華航空
Heat wave
猛暑
Democratic Party of Japan
民主党
Pension
国民年金
most splogs from professional spammers
COMSN, Inc.
コムスン
Asashōryū
朝青龍
Duration: Long Term
Duration: Short Term
The dignity of the woman
女性の品格
Darvīsh
ダルビッシュ
Health food
健康食品
ZARD
Cosmetic surgery
美容整形
Saeko
サエコ
Lazy woman
干物女
most splogs from professional spammers
Billy's Boot Camp
ビリーズブートキャンプ
Viagra
バイアグラ
Erog
エログ
Wii
Rumor
ウワサ
Video
動画
No revision
無修正
Private Concern
・red keywords: splog rate > 10%
・with box: splog rate > 30%

Splog Rate, Professional Spammer Rate, Amateur Only Splog Rate

Kewords	Splog rate (splog / (splog + non-splog))	Professional spammer rate (from professional spammer / splog)	Professional Spammer ID	Amateur Only splog rate (from amateur spammer / (from amateur spammer + non-splog))
エログ (Erog, adult)	89.2%	**92.4%**	**2, 6, 8**	38.5%
ウワサ (rumor)	88.1%	**94.8%**	**1**	27.8%
国民年金 (national pension)	58.1%	**90.2%**	**3, 4**	12.0%
無修正 (no revision)	40.9%	18.5%	**1**	36.1%
健康食品 (health food)	37.4%	**58.7%**	**5, 7**	19.8%
美容整形 (comsmetic surgery)			**10**	21.7%
バイアグラ (Viagra)				20.5%
ダルビッシュ (Darvish, baseball player)			-	22.1%
動画 (video)			-	19.1%
朝青龍 (Asasho-ryu, sumo wrestler)	15.2%	**80.0%**	**1, 6**	3.4%
ビリーズブートキャンプ (Billy's Boot Camp)	15.1%	0.0%	**---**	15.1%
サエコ (Saeko, Darvish's wife)	14.3%	14.3%	**1**	12.2%
Total of 22 keywords	23.5%	**61.5%**	**10 Groups**	**9.0%**

Keywords with high amateur only splog rates are those with private concern.


Public Concern
North Korea
北朝鮮
Upper House election
参議院選挙
China Airlines
中華航空
Heat wave
猛暑
Democratic Party of Japan
民主党
Pension
国民年金
Duration: Long Term
COMSN, Inc.
コムスン
Asashōryū
朝青龍
The dignity of the woman
女性の品格
Darvīsh
ダルビッシュ
Duration: Short Term
Health food
健康食品
ZARD
Cosmetic surgery
美容整形
Saeko
サエコ
Lazy woman
干物女
Billy's Boot Camp
ビリーズブートキャンプ
Viagra
バイアグラ
Erog
エログ
Wii
・red keywords: splog rate > 10%
・with box: splog rate > 30%
Video
動画
No revision
無修正
Rumor
ウワサ
Private Concern


Public Concern
North Korea
北朝鮮
Upper House election
参議院選挙
China Airlines
Heat wave
猛暑
Keywords with
high amateur only splog rates are
those with private concern.
COMSN, Inc.
コムスン
Asashoryu
朝青龍
Duration: Long Term
Duration: Short Term
The dignity of
the woman
女性の品格
Darvish
ダルビッシュ
Health food
健康食品
Cosmetic surgery
美容整形
ZARD
Saeko
サエコ
Lazy woman
干物女
Billy's Boot Camp
ビリーズブートキャンプ
Viagra
バイアグラ
Wii
Erog
エログ
Rumor
ウワサ
Video
動画
No revision
無修正
Private Concern
・red keywords: splog rate > 10%
・with box: splog rate > 30%

Manual Analysis of Splogs: Summary

- Blog Host Distribution･･･Top-3 hosts cover 88% of Splogs
- Classifying Spammers into *Professional / Amateur*
  - Analyzing Professional Spammers･･･10 Pro Spammers
- Splog Features and Keywords
- Correlation:
  - Keywords with Public Concern have Low Splog Rate.
  - Keywords with Private Concern may have High Rate of Amateur Only Splogs.

Future Works

- Integration with previously studied features of splogs:
  - characteristic words in splogs,
  - in-degree/out-degree distributions,
  - ping time series.

- Applying existing machine learning based splog detection techniques, and developing a splog detector with high accuracy.

- Semi-supervised framework of detecting splogs of previously unobserved types.

Thanks for your attention!