

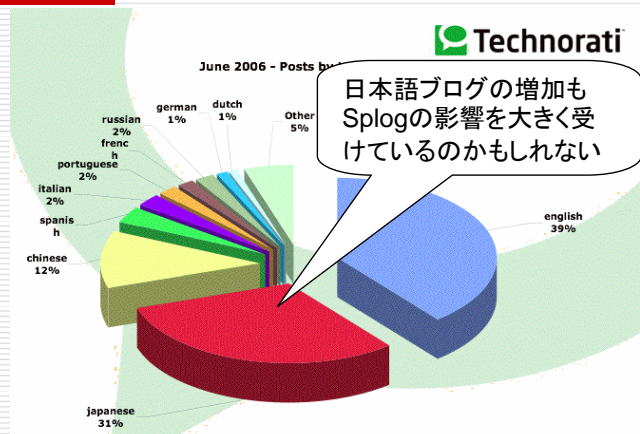
スパムブログ (Splog) に関する 調査報告

- 東京大学人工物工学研究センター・福原知宏
- 東京大学・中川裕志教授
- 筑波大学・宇津呂武仁准教授, 佐藤有紀君 (M1)
- 東京電機大学・増田英孝准教授, 芳中隆幸君 (B4)
- (株)ナビックスとの共同研究

2007年11月4日(日)
科研情報爆発・NLPミーティングにて発表

背景

- スпамブログ (Splog) の増加
 - 当初は英語圏に見られたが、近年は英語圏以外でも増加
- 経緯
 - 2005年: Bloggerで問題になる
 - 2006年: Technoratiのレポート
 - 日本語ブログの量は世界2位となったが、Splogも相当混じっていると考えられる。



スパログ【splog】とは
(デイリー新語辞典(三省堂)より)

インターネットにおいて、リンク誘導の目的で大量に自動生成される、内容には意味のないブログのこと。目的サイトへのリンクを多数埋め込んだエントリー(記事)が大量に生成されるため、検索エンジンの表示順位が不当に操作されるなどの不具合が生じる。

ブログ空間における使用言語の分布(2006年6月)
<http://www.sifry.com/alerts/archives/000436.html>

ある日のSplog状況

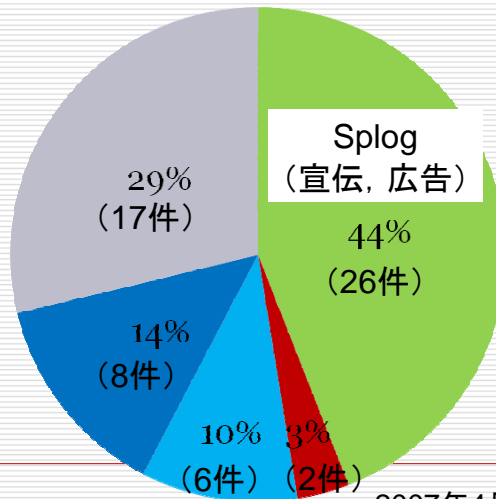
1日に10件以上記事を投稿しているブログサイトのうち、Splogの割合

Splog割合

■ 宣伝、広告 ■ 画像 ■ ニュース ■ 日記 ■ なし

全59サイト中

宣伝、広告	26件
画像	2件
ニュース	6件
日記	8件
なし	17件



およそ半数(44%)がSplog

2007年4月30日のブログ収集データを元に
東京電機大学・芳中隆幸君調べ

背景： 何故ブログでスパムか？

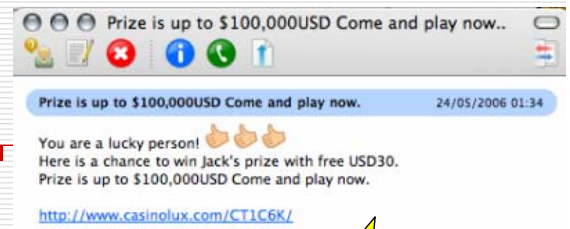
- ブログ(CMS)とブログサービスの登場による情報発信コストの低下
 - ← それまではレンタルサーバ等で、手作業でWebページ作成しなければならなかった
- アフィリエイトプログラムの登場によるビジネスチャンスの拡大

→個人がスパマーに変貌する機会の増加



仁義無きスパム

- スパムメール
- 掲示板スパム(掲示板荒らし)
- リンクスパム(スパムページ)
- スパムブログ
- Comment/trackbackスパム
- Mixi あしあとスパム
- Skype スパム
- Twitter spam
- YouTube スパム
- Wikipedia スパム
- ソーシャルブックマークスパム
 - Del.icio.us Spam
 - はてブクラッシャー
- オンラインゲームスパム(Smog; spam in multiplayer online games)
 - Second Life スパム
- etc...

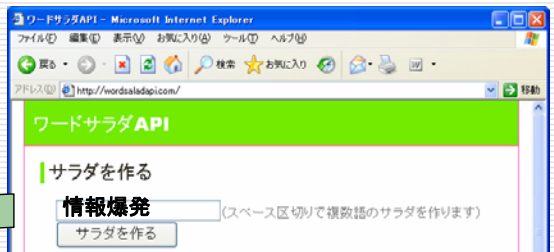


Twitter spam

Spamは至る所に存在する

Splogの種類

- 自動生成型(ワードサラダ型)
 - 予め用意した辞書から単語を無作為に取り出し意味を成さない文章を自動生成するタイプ
 - 例:ワードサラダAPI
- コピペ型
 - ニュース記事や他人のブログ記事を勝手に引用するタイプ。
 - コピペしたテキストにコメントを書き足す場合もある。
 - 例:ウワサスパム
- リンクファーム型
 - 相互にリンクし合うことでページランクを向上
 - 例: Keyword Vampire, アダルトブログ



ワードサラダAPI
http://wordsaladapi.com/

情報爆発の本質

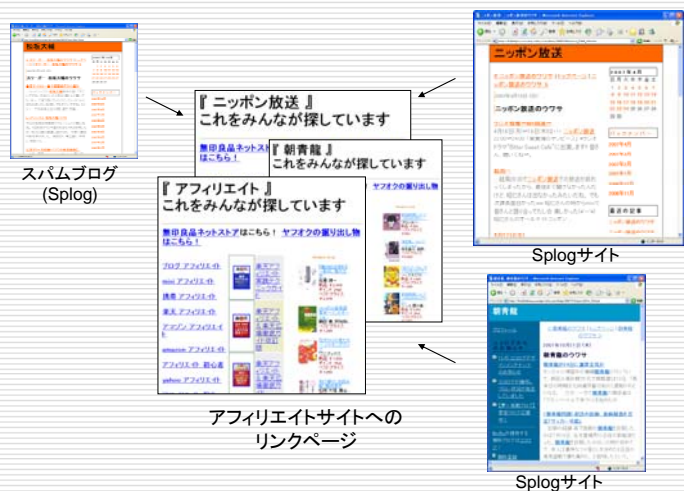
そのうちのひとつが江原啓之の“素人スピリチュアル・カウンセリング企画”だった。口が焼けるような感じがするのは。サービスの“キモ”であるコメント機能の作り方を明かした。



コピペ型
(ウワサスパム)

Splog事例: ウワサスパム

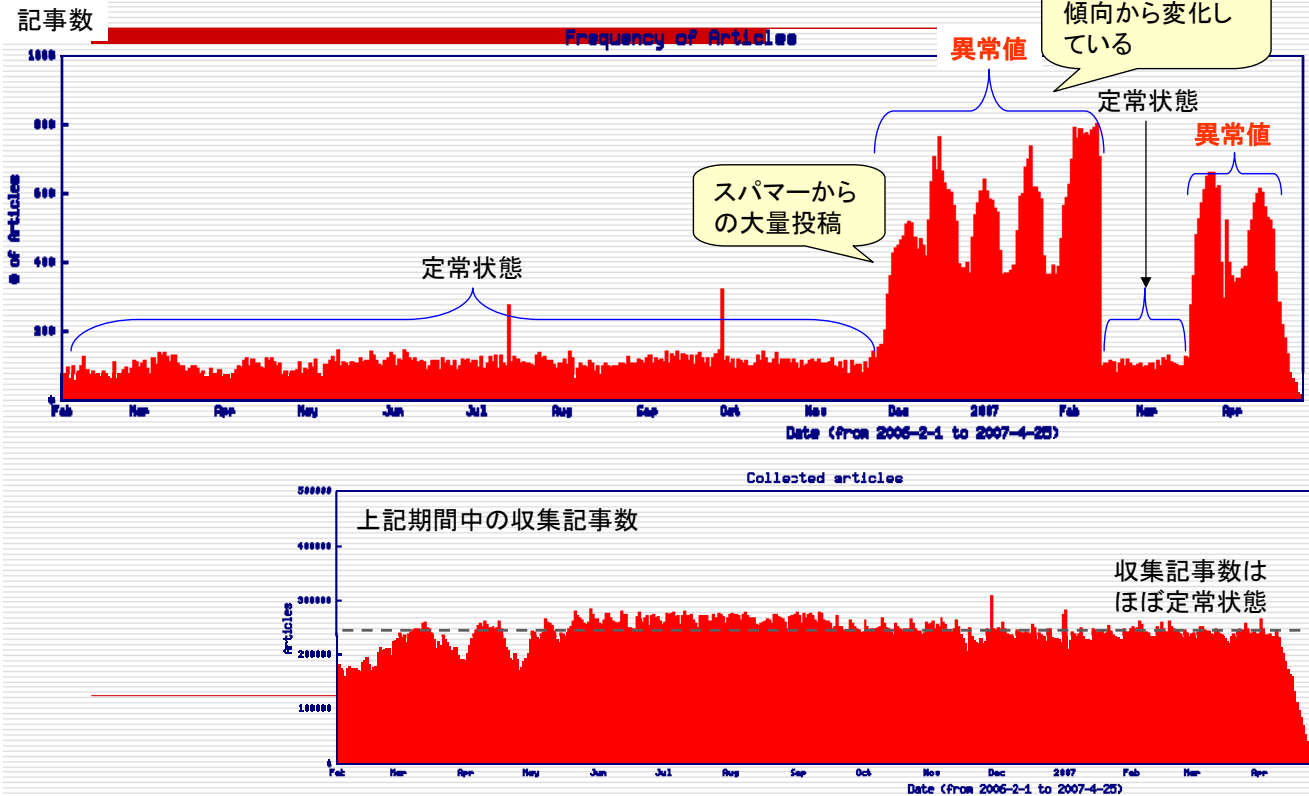
- 「〇〇のウワサ」というタイトルで大量投稿
 - 例: http://kj5e6jss.cocolog-nifty.com/blog/2007/04/post_27b8_14.html
※ホスト名: kj5e6jssも辞書に無い無意味な文字列
 - 記事の中身は他のブログ記事からコピー
 - h-itc.net(アフィリエイトサイト)にリンク
 - アフィリエイトサイトからYahoo!ショッピング, ヤフオク等にリンク



ウワサスパム(続き)

This section provides a detailed look at the spamming technique. On the left, a screenshot of a splog site shows a post titled '『ニッポン放送』これをみんなが探しています'. A callout box points to the text 'ヤフオク 無印良品でのアフィリエイト稼ぎ' (Earning affiliate commissions from Yahoo! Auctions and Muji). On the right, a screenshot of an affiliate site shows a post titled 'ニッポン放送'. A callout box points to the text 'キーワードごとにブログを開設' (Creating a blog for each keyword). Another callout box points to a link '「ニッポン放送のウワサ」トップページ' with the note 'クリックするとアフィリエイト稼ぎのページに飛ぶ' (Clicking leads to an affiliate earning page). A third callout box points to a list of dates (2007年2月, 2007年1月, 2007年10月) with the note '他人のブログ記事のコピペ' (Copying others' blog posts). A final callout box points to a list of recent posts with the note '他人のブログ記事のコピペ' (Copying others' blog posts).

語の出現頻度： “ウワサ”の出現頻度推移



Splogの特徴

□ 自動生成ブログ

- ★ サテライトサイトの自動生成ツール(例:Keyword Vampire)
- ★ 日々更新され、投稿時間が一定

□ アダルト系ブログ

- ★ 訪問者獲得のため、芸能人の名前を羅列

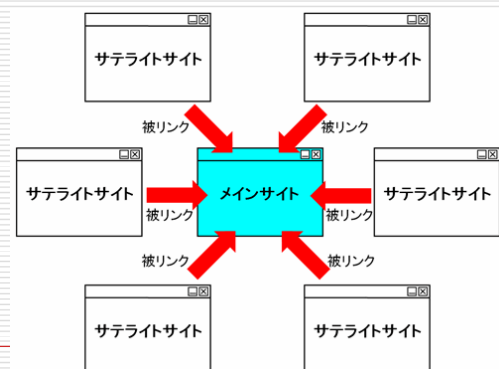
□ 商用ブログ

- ★ トップページに大量のリンク情報を掲載



Keyword Vampire Pro
(9,870円, 今だけ特別価格)

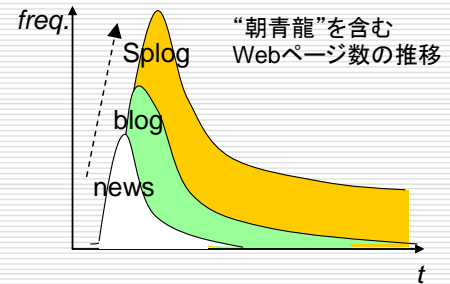
あたかも吸血鬼(ヴァンパイア)が咬みついて仲間を増やしていくかのように、サイトを増殖させることができるので、ソフト名も「キーワードヴァンパイアプロ(Keyword Vampire Pro)」です！！



Keyword Vampire Pro ホームページから

日本語ブログ分析の主な知見

- ブログサービス各社による対応の違い
 - ブログサービス各社によってスパム率は異なる
 - Yahoo!はスパムをきっちり排除
- 日本語ブログでは、バーストに便乗してスパムが発生する傾向
 - その時々話題を表すキーワードと結びつく(朝青龍, 時津風部屋, 亀田兄弟...)
 - 訃報トピックに対してはスパム率低い(e.g., ZARD)
 - バーストした日よりも、その直後にスパムが大量発生(時間差を伴う)

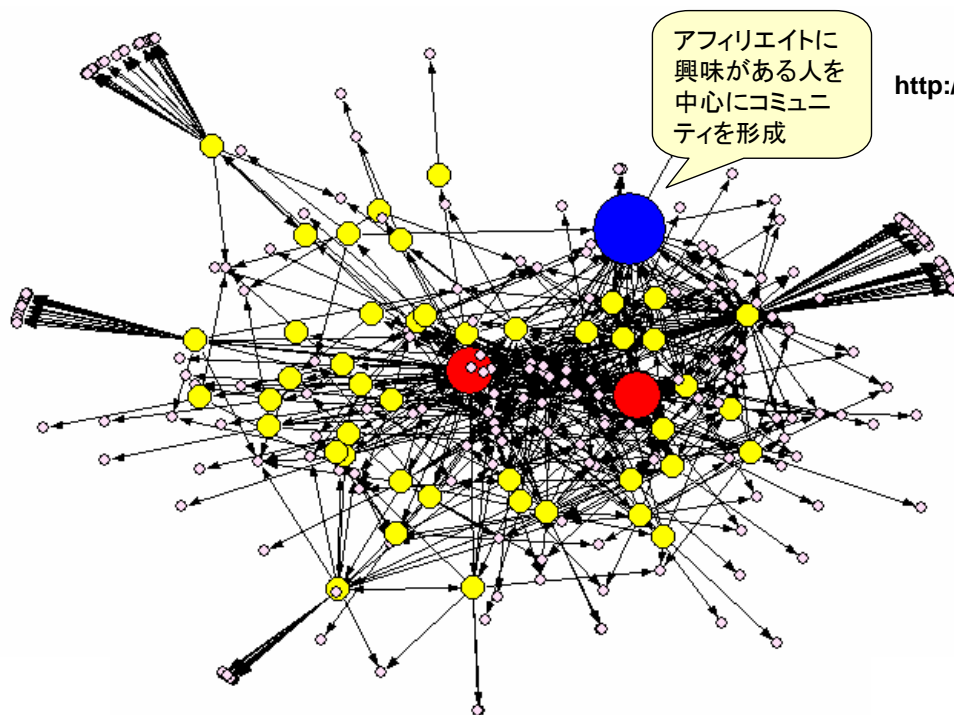


news, blog, Splogの関係(仮説):
時間差を伴って広がっていく?

ブログ間ネットワークを用いたスパム検出 (調査中)

- ブログ空間におけるネットワーク構造を知る
 - スパムブログはネットワークを形成
 - Webページをクローリングしてネットワーク構造を抽出. Pajekでグラフを出力.
 - アフィリエイト, アダルト, 消費者金融に関するブログからネットワークを抽出

ブログ間ネットワーク (情報商材・アフィリエイト)

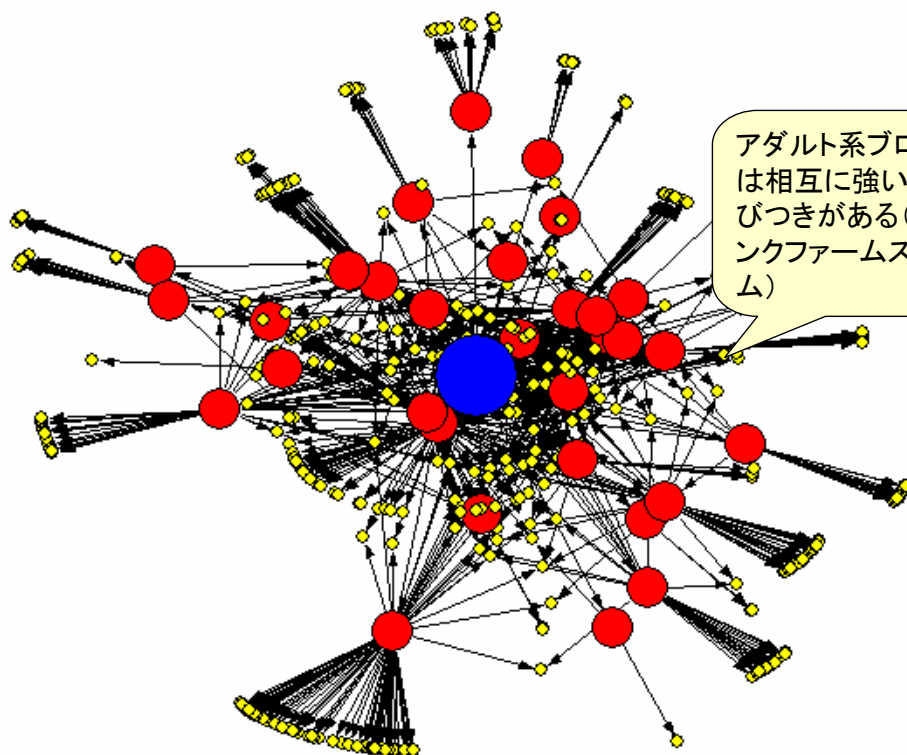


アフィリエイトに興味がある人を中心にコミュニティを形成

FX情報商材マキシマム
<http://fxsysteminfo.blog75.fc2.com/>

- 開始ページ (第1階層)
- 第2層
- 第3層

ブログ間ネットワーク (アダルト)



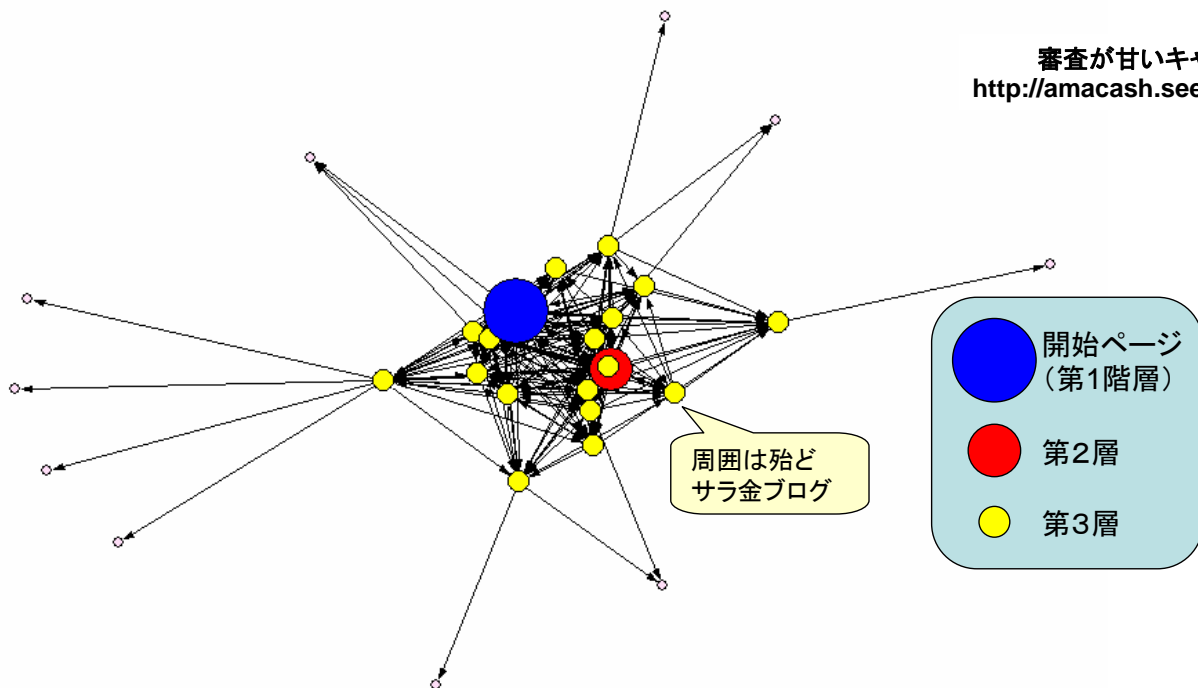
アダルト系ブログは相互に強い結びつきがある(リンクファームスパム)

エログちゃんねるにゅーす
<http://www.elog-ch.com/news/>

- 開始ページ (第1階層)
- 第2層
- 第3層

ブログ間ネットワーク (消費者金融ネットワーク)

審査が甘いキャッシング
<http://amacash.seesaa.net/>



Splogまとめ

- Splogと格闘しなければならない時期に来ている.
 - Splogの発生には傾向が見られる.
 - キーワードがバーストした後にSplogが増加.
 - Splogとネットワークの構造については調査中
 - 中心と関連するブログが周囲にある.
 - Fake blog (FLOG; やらせブログ)はSplogとは異なるが読者を困惑させる点でスパムと見なせる.
 - Splog, FLOGとも最後は「信用・信頼」が鍵となる?
- 現在の作業
 - Splog評価用データセットの作成(日韓ブログ)
 - Splogフィルタの開発(筑波大宇津呂研にて)