3E2-4

# インターネットセキュリティとリスクマイニング

Risk Mining on Internet Data

吉田 健一
Kenichi Yoshida

筑波大学 大学院 ビジネス科学研究科
Graduate School of Business Science, University of Tsukuba

現在 2011 年の地上波デジタル放送への完全以降・インターネットを使った放送事業の胎動や、インターネットを使った各種販促技術の急速な発展にともない、年間２兆円の市場規模を持つ TV 広告事業・放送業界の事業形態が変容しようとしている。中でも今までは人手による視聴率調査以外に有効な方法のなかった TV 広告の効果計測が、インターネット掲示板や blog の解析により、実時間で、且つ、個々の出演者に対する好感度調査なども同時に実施する形で行える可能性が生じつつある事は社会的影響が大きいことが予想される。

このように情報技術が新しい価値を創出しようとしている一方、迷惑メイルやインターネットウイルス等のマイナス面が新なマイナスの社会要因を作りつつある。本発表では、このような社会背景から望まれるリスクマイニングの諸側面について討議を試みる。

## 1. はじめに

現在 2011 年の地上波デジタル放送への完全以降・インターネットを使った放送事業の胎動や、インターネットを使った各種販促技術（WWW を使った顧客管理など) の急速な発展にともない、年間２兆円の市場規模を持つ TV 広告事業・放送業界の事業形態が変容しようとしている。中でも今までは人手による視聴率調査以外に有効な方法のなかった TV 広告の効果計測が、インターネット掲示板や blog の解析により、実時間で、且つ、個々の出演者に対する好感度調査なども同時に実施する形で行える可能性が生じつつある事は社会的影響が大きいことが予想される。

この背景には、インターネットを利用する人々が国民の約６割強に達し、また消費行動の中で重要な位置を占めつつある事とともに、インターネット上の各種データを解析する情報技術の進展が重要な意味を持っている。

このように情報技術が新しい価値を創出しようとしている一方、迷惑メイルやインターネットウイルス等のマイナス面が新なマイナスの社会要因を作りつつある。本発表では、このような社会背景から望まれるリスクマイニングの諸側面について討議を試みる。

## 2. インターネット・マーケティングの進展

インターネットの浸透がマーケティングに及ぼす好例として、2ch と呼ばれるインターネット掲示板と、そこに投稿される記事の内容を分析する事による TV 視聴率の実時間解析技術がある。

上原は実況スレッドと呼ばれる 2ch の記事が、放送中の人気ドラマ番組に対して、１時間あたり数千という膨大な数になる事に着目し、キーワードの単純な出現回数計測だけで、特定の俳優への分単位での視聴者の興味の変化が分析できる事を報告している [1]。例えば図 1 は、掲示板から抽出した出演者に対する１分ごとの着目度の変化（黒い線）と、アンケート調査による着目度の変化（灰色の線）が一致している様子の示している。図から明らかなように掲示板から簡単に取れる情報は、今まで入手するにはアンケートの実施など人手が必要であった情報と精度良く一致する。TV 広告が産業に及ぼす影響を考えた場合、この事の応用価値・社会的意義は大きい。

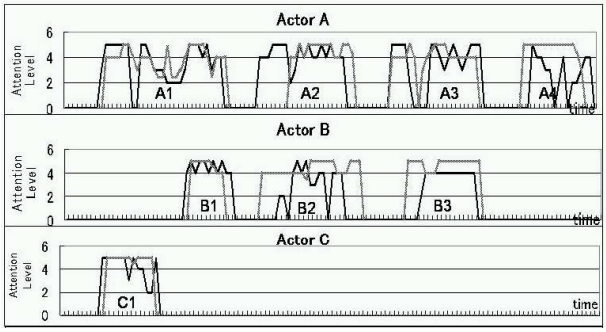


図 1: 2ch の解析による TV ドラマ視聴の分析

このような解析を可能にした背景には、若者を中心に、2ch に代表されるインターネット掲示板を見ながら TV を視聴し、視聴と同時に TV 番組に対するコメントを掲示板に投稿する TV 視聴者が増えており、結果として、そのようなインターネット掲示板の情報量が増えており、従来のような複雑な自然言語解析技術を用いなくても単純な統計的な処理だけで、番組出演者に対する注目度を抽出することができるようになってきた事がある。

この事は

- 人手をかけずに分単位で番組出演者に対する視聴者の反応の大きさが実時間で計測できる。
- 上記は、単純な視聴率調査にとどまらず、分単位でどの出演者に着目が集まっているかという、時間単位においても、解析単位（この場合は出演者）においても詳細な情報が解析可能である。
- 2ch は日本特有のインターネット掲示板であるが、blog など、類似の WWW 技術を用いた、TV とインターネット掲示板の同時視聴の形態は、若い人を中心に世界的に広がる可能性がある。この場合、関連する広告業界などへの社会的影響は世界的に見ても大きい。

連絡先: 〒 112-0012 東京都文京区大塚 3-29-1
yoshida@gssm.otsuka.tsukuba.ac.jp

と言った観点から関連研究分野への大きな影響が予想される。

また、同様な研究は、スポーツ番組を対象とした宮森の研究 [2] や、ニュース番組を対象とした上原の研究 [3] など事例を増やしつつあり、インターネットの社会への浸透を背景に、新しいマーケティング技術として重要性を増しつつある。

これらは、単純なデータマイニング技術であっても、インターネットの持つ強力な情報収集能力と組合せる事により、インターネット時代の商業広告をささえる情報技術／マーケティング技術になりえる事を示唆している。

## 3. インターネット・セキュリティの課題

上記のようなインターネット時代の商業広告をささえる情報技術／マーケティング技術の研究が進む一方、広告に関連して、行き過ぎた技術利用による迷惑メイルが社会問題化している。道徳的にゆるされるか否かの判断は別にして、迷惑メイルを広告技術としてみた場合

- 名簿作成専業業者が存在するなど、潜在顧客リストの自動作成の仕組が確立している。
- 安価に宣伝文を配布可能である。
- 反応してきた有望顧客リストの自動作成と名簿管理も、自動化されている。
- 高い利益率と速い反応を備えている。

と非常に優れた性質を持っている。この事は、もともと研究者間の実験ネットワークとして設計された経緯を持つ、正善説に基くインフラ（TCP/IP）技術と、やはり研究者間の連絡のために設計されたアプリである WWW/mail の特質とも重なり、迷惑メイルの対策が進まないことの一因にもなっている。

また飽くなき利便性の追及はシステムの複雑化を招き、セキュリティホールを増やす結果に繋がっている。例えば商用／非商用を問わず、使いがっての良い WWW ブラウザーにセキュリティホールが見つかるという事例は良く発生する。新種の Internet Virus も日常的に報告・警告されている。

## 4. リスクマイニングへの期待

前述のような迷惑メイルや Internet Virus の中にはデータマイニング技術を応用する事により、比較的容易に対策できるものもある。例えば、迷惑メイルに関して「類似したメイルの数を数え一定数以上のものをスパムと分類する」という単純な頻出 item 検出のアイデアが有効な事が実証的に示されている [9]。同様なアイデアはインターネットセキュリティの分野では各所に使用できると思われる。例えば Internet Virus も頻出する src IP adress と destination port number の組み合わせをバックボーンに流れるパケットの情報から抽出する事で検出可能に思われるし (図 2)、分散 DoS 攻撃の検出は特定の計算機に多数の計算機から多量の syn パケットが送られている事を検知できれば検出可能に思われる。セキュリティ向上の観点から P2P によるネットワークの使用状況を把握しておく事はネットワークの重要な管理業務の一部であるが、これも特定の src IP adress と src port number の組み合わせとして検出可能に思われる。

これらは何れも単純な頻出 item または itemset の検出技術が迷惑メイルやインターネットウイルス等のマイナス面など新なマイナスの社会要因への対応技術として有望な事を示唆し、リスクマイニングへの期待となっている。

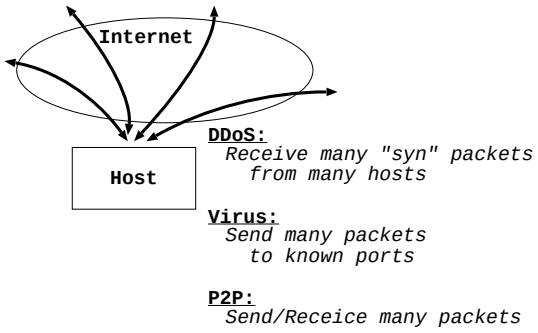


図 2: Virus/DDoS/P2P の検出

## 5. リスクマイニングの研究課題

前述のように、迷惑メイルやインターネットウイルス等の新たなマイナスの社会要因を防ぐ手段として、データマイニングの技術は有望と考えるが、いくつかの技術的なハードルが残されている。例えば、迷惑メイルやインターネットウイルス等の対策を考えた時には、大量のオンラインデータからリアルタイムでデータマイニング [5] を行おうとする試み [6] の重要性が高まっている。特に最近の研究で、WWW や spam mail など Zipf の法則 [7] に従うデータが多い事がわかってきた (例えば [8], [9]) が、そのような性質を持つ大量のデータからの規則性抽出は重要なテーマである。

図 3 に [10] で取り上げられている Internet backbone のパケット分布を示す。このようなデータからマイニングを行うことは前述のように Internet Virus や分散 DoS 攻撃を検出する事になり、社会リスク低減を目的とした技術としての価値が高いが、このようなデータは非常に大量であり、高速に生成される。従って、データに比較すれば少量のメモリで高速に動くマイニング技術開発の重葉性が増すと考えるが、上記データは単純な Zipf の法則に従っているわけではなく、手法の開発には、そのようなデータの特性自体の研究も重要と考える。

例えば、図 4 に図 3 に示したデータの積算値をプロットした図を示すが、単純な Zipf の法則に従ったデータ (図で実線) と実データ (図で点線) は大きく異っている。

データが図 3,4 に示すような分布を示した場合、1Gbps のアクセス回線に流れるネットワークパケットのデータ量は、概ね、表 1 に示したようになるが、これは代表的な頻出 itemset 抽出プログラム LCM-v2[11] で解析を試みた場合、非常に大きなメモリ用量を必要とする事を意味する。図 5 に前述の Internet backbone のパケットデータを LCM-v2 で解析した時の、パケット数とプロセスサイズの関係を示す。図から明かなように、LCM-v2 は 1Gbps のアクセス回線に 40 秒間に流れるネットワークパケットを解析するのに 2Gbyte 弱のメモリを必要とする。

ネットワークの保守管理にはこの種の解析を 1 時間単位で実施する事が求められる事が普通であり、現状の頻出 itemset 抽出プログラムでは事実上解析する事はできない。

## 6. リスクマイニングへのアプローチ

著者らは大量のオンラインデータから小容量のメモリを使ってリアルタイムで頻出するアイテムの組合せを抽出する方法として、固定サイズのキャッシュを使う方法を検討している [10]。「頻出するアイテムの組合せ」ではなく、「頻出するアイ

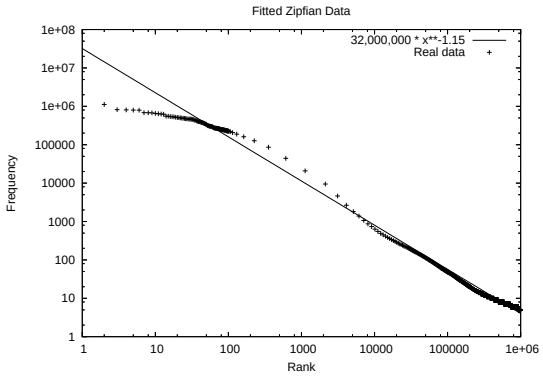


図 3: ランク順に並べたデータの出現頻度

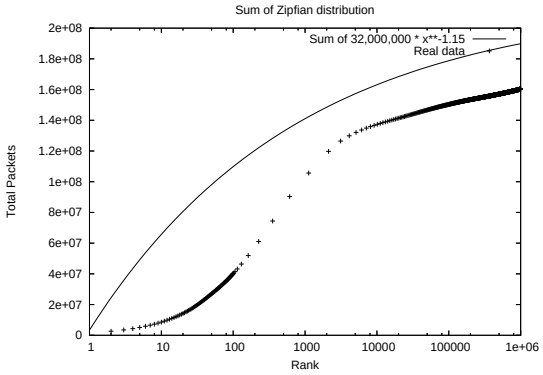


図 4: ランクまでのデータ出現頻度の合計

表 1: ネットワークパケットのデータ量

| | Number of Packets | Kinds of Packets |
|---|---|---|
| 1 second | 1M | 1K |
| 1 minute | 60M | 4M |
| 1 hour | 4G | 210M |
| 1 day | 86G | 4G |
| 1 week | 605G | 35G |

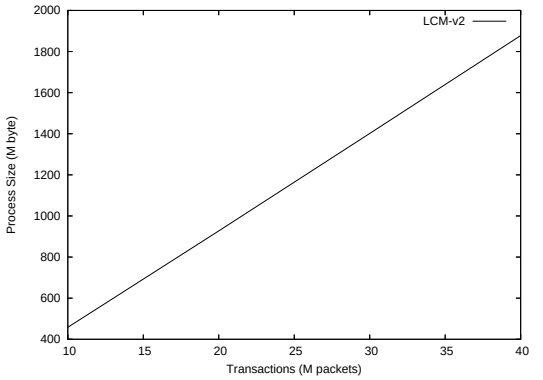


図 5: LCM-v2 による解析に必要なメモリ容量

```
Create empty heap;
while (input item) do
  i = index of item in heap;
  increment heap_cnt[i] by 1;
  if (heap_cnt[i]>thresh_hold)
    print message;
done
```

図 6: Cache-based Pattern Mining Algorithm

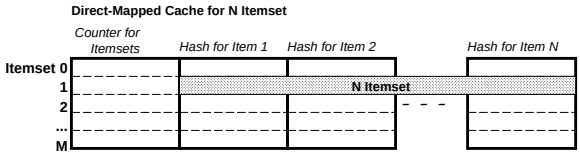


図 7: Cache-based Pattern Mining Data Structure

テム」の抽出に限定すればアイデアは単純であり、固定サイズのキャッシュを使って処理対象のデータに含まれるアイテムの数を数え、設定値以上の回数出現したアイテムを頻出アイテムとして出力する (図 6 に疑似コードを示す)。

頻出するアイテムの組合せを検出する場合、キャッシュの構造が若干複雑になる (図 7) が、基本的には同じ考え方が使える。すなわち、初期は頻出アイテムの抽出を行っておき、頻出すると判断されるアイテムが見つかった後はその頻出するアイテムを含む 2 つのアイテムの組合せについても出現回数をチェックする。N 個のアイテムの組合せが頻出すると判断された後は、その N 個のアイテムの組合せを含む N+1 個のアイテムの出現回数をチェックする。正確には処理速度をあげるための工夫などが必要となるが、ここでは詳細は割愛する。

キャッシュの容量が十分大きい場合、アイテムの記憶位置を決める処理 (図 6 の 3 行目) は単純である (過去に出現して記憶されているものは hash 関数などを用いて記憶位置を決定し、新しいアイテムであれば空きエリアを記憶位置とする) が、容量が全アイテムを記憶するだけ確保できない場合、なんらかの指標に基づき既にあるアイテムのデータを削除して、キャッシュメモリを再利用する事を考える。

LRU は、このような時に用いられる標準的な方法であるが、実験では LRU に比べて単純な random2(図 8) の性能が良いことがわかってきた [10]。random2 は [9] にて SPAM filter を作る時のキャッシュ管理の方式として優れた性能を持つ事が報告されている方法で、図 8 は C による random2 のプログラムコードそのものである。1 行目は random そのもの、2 行目と 3 行目は複数回出現したデータをなるべく捨てないようにしている処理を実現しているが、基本的にはこの修正をした後も、乱数で決めた数をキャッシュサイズで割って余りが新しいアイテムの記憶位置 (削除するデータの記憶位置) となっている。

実際にネットワークのデータから頻出 itemset を抽出するに

```
int i = random() % HEAP;
for (p=1; (heap_cnt[i]>p); p++)
   i = random() % HEAP;
```

図 8: Random2

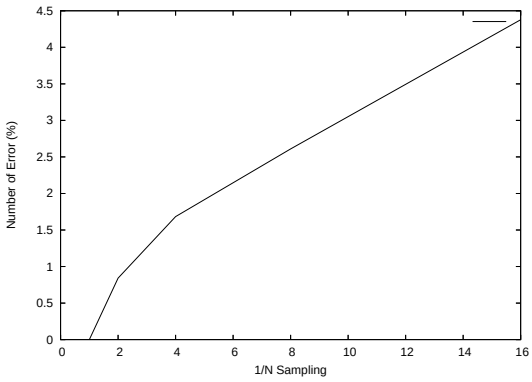


図 9: サンプリングエラー

は処理時間の問題もあるため、サンプリング手法を組合せて解析する必要がある。この場合、サンプリングの影響で、頻出 itemset を見逃したり、頻出 itemset でないものを頻出 itemset と誤認するようなエラーが問題となるが、提案手法は 1/16 程度まで許容可能なエラーでデータが解析可能であり (図 9)、メモリ容量、処理時間、精度の観点から、実時間でデータ解析が可能な方法になっている。

著者らの検討している手法は単純ではあるが、図 3,4 に示したデータを解析する上では優れた特色を持っており、今後応用面まで含めた研究を進める予定である。

## 7. まとめ

新しい情報技術が新しい価値を創出しようとしている一方、迷惑メイルやインターネットウイルス等のマイナス面が新なマイナスの社会要因を作りつつある。本発表では、このような社会背景から望まれるリスクマイニングの諸側面について討議を試み、

- 新しい情報技術が社会に及ぼしつつあるプラスの影響と、
- 迷惑メイルやインターネットウイルス等のマイナス面、
- 著者らが研究している上記マイナス面に対する技術的な課題、

について述べた。

本報では主に技術的側面について取り上げたが、新しい情報技術が持つマイナス面を対策するには、技術的な検討だけでは不十分である。例えば通信内容のマイニング処理は盗聴に繋がる側面を持っており、使用に際してはリスク対策が別の社会的に見た時に負の要因を生成しないか議論した上で使っていく必要がある。どこまでの解析を社会的に容認するかについては 1 つの技術の正の側面・負の側面、両面から見た上で社会の中でその功罪について議論し合意を形成していく必要がある。現状、技術的な開発に社会的な合意形成が追い付いていっていない事も大きな問題となっていると考える。

## 参考文献

[1] Hiroshi Uehara, Kenichi Yoshida, “Scripting TV Drama based on Viewer Dialogue – Analysis of Viewers’ Attention Generated on an Internet Bulletin Board –,” In Proc. of SAINT, pp 334-340 14.

[2] 宮森　恒・中村聡史・田中克己, “番組実況チャットを利用した放送コンテンツの自動インデキシング、” 電子情報通信学会パターン認識・メディア理解研究会 予稿、NLC2004-123 PRMU2004-205, pp.43-48

[3] 上原　宏, “携帯端末とネットコミュニティ連携の可能性,” 平成１７年度 情報化月間行事 ＩＴシンポジウム

[4] 松村真宏，三浦麻子，柴内康文，大澤幸生，石塚満、“２ちゃんねるが盛り上がるダイナミズム，” 情報処理学会 45 巻 3 号, pp. 1053–1061, 2004.

[5] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” Proc. 20th Int. Conf. Very Large Data Bases, VLDB, ed. J.B. Bocca, M. Jarke, and C. Zaniolo, pp.487–499, Morgan Kaufmann, 12–15 1994.

[6] C. Hidber, “Online association rule mining,” pp.145–156, 1999.

[7] 水谷, 数理言語学, 培風館, 1982.

[8] N. Nishikawa, T. Y.Mori, K.Yoshida, and H.Tsuji, “Memory-based architecture for distributed www caching proxy,” In Proc. of World Wide Web Conference 98, pp.205–214, 1998.

[9] Kenichi YOSHIDA, Fuminori ADACHI, Takashi WASHIO, Hiroshi MOTODA, Teruaki HOMMA, Akihiro NAKASHIMA, Hiromitsu FUJIKAWA, Katsuyuki YAMAZAKI, “Density-based spam detector,” 電子情報通信学会 英文誌 D 特集号, Vol. E87-D, No12, 2004

[10] 吉田健一, 勝野聡, 藤田昌克, 鶴正人, 阿野茂浩, 山崎克之, “キャッシュを使った頻出アイテムの抽出,” 電子情報通信学会　和文論文誌 10 月号 (B 分冊 12) Vol.J88-B No.10 p.2012

[11] http://sunsite.informatik.rwth-aachen.de/publications/ceur-ws//vol-126/.