サイエンティフィック・システム研究会 2015 年 10 月 29 日
合同分科会 2015 年度会合 資料

ネットワークデータの異なり数計測とその応用

吉田 健一

筑波大学 教授

[キーワード]
ネットワーク計測、異なり数、データマイニング、セキュリティ、社会科学

1. はじめに

幾つの計算機の間で通信が行われているか解析する異なり数の計測技術は、概念が簡単である一方、応用範囲が極めて広い。本講演では、その基本的な概念(一般的な延数との違い)から、DDoS, Internet Virus などセキュリティインシデントの検出、ネットマーケティングの効果測定・社会調査への利用まで、応用事例を紹介する。

2. 異なり数解析とは [1]

「延べ数」という言葉は良く用いられる。何らかのデータの集計も延べ数で基礎となる数字を計測している事が多いように思う。「異なり数」とはその対義語である。今、ある WWW サーバに A 氏から H 氏まで 8 人が 1 回ずつアクセスした場合、延べ数で数えれば 8 回である。英語論文で「total number of access」と記されていたら通常はこちらの「延べ数」を意味している。一方ある WWW サーバに X 氏と Y 氏が 4 回ずつアクセスした場合、延べ数で数えれば同じ 8 回であるが、異なり数で解析しようとした場合、「2 名からアクセスがあった」と数える。

両例は WWW サーバから見たアクセス負荷は同じ(同じ 8 回分のアクセス負荷)であり、ネットワークトラフィックの研究では延べ数を使った計測結果が利用される事が多いように思う。しかし社会科学の解析として、例えば「熱狂的なファンが 2 名いてアクセスが 8 回になった」のと「8 人が 1 回ずつアクセスした」のでは意味が異なる。熱狂的少数のファンを対象にビジネスを考えるか、広く受け入れられる商品でビジネスを考えるか、計測した「数」の利用方法にも関係してくる。どちらで数を数えるべきか自体が検討テーマになりえる。

3. 不正侵入/Virus/DDoS の発見 [2]

今日ではインターネットの運用管理は社会インフラ維持の観点から重要な研究テーマである。特にセキュリティ面での監視は社会からのニーズも高い。従来からインターネットの運用管理の基礎情報として用いられてきた帯域監視は、回線の利用状況や、アプリケーション毎の利用比率は延べ数による数値データであり、運用管理や将来の設備計画の基礎情報として欠くことのできない情報である。しかしながら、このデータからは不正侵入や Internet Virus、DDoS 攻撃といったセキュリティ面での情報は得られない。例えば DDoS 攻撃は攻撃目標のサーバに多量のパケットを送って動作不全に陥らせる攻撃手法を用いるが、多量と言っても「1 サーバの受け取るパケットとしては多量」と言う事であり、帯域のような情報だけから検出する事は難しい。インターネット全体で見れば通常の WWW 等のトラフィックの方が遥かに大量であり、その影に隠れてしまう。

異なり数と言うアイデアは、このような場合に有効な監視手段を与えてくれる。具体的には延べ数だけでなく、異なり数も大きなパケットを探せば、異常なトラフィックとしてセキュリティに関るパケットを検出できる事が多い。例えば DDoS 攻撃であれば、攻撃目標となったサーバは多数の計算機からパケットを受け取るので、送信元 IP

アドレスの異なり数が大きな、宛先 IP アドレスを監視していれば、検出できる。Internet Virus に感染した計算機は次に侵入する計算機を探すために多くの計算機に通信(すなわち侵入の試み)を行う。Internet Virus は次々と新種が生れるが、多くの計算機に通信を行うという特徴は共通であり、宛先 IP アドレスの大きな送信元 IP アドレスを監視していれば、怪しげな計算機を検出できる。

4. 広告宣伝の効果分析 [3]

携帯電話を端末としたインターネット経由の情報交換が一般的になっており、「ビッグデータ」やら「パーソナルデータ」は広告等のビジネスを進める上での重要なバズワードになりつつある。旧来のマスメディアからインターネットに投資が移っている現象の背景は、ネットの情報拡散能力が高い事だけではない。ユーザーが WWW 上の広告をクリックし、リンクされたページを閲覧した時点で初めて広告料金が発生するクリック課金型広告など、効果が直接観測できる宣伝形態が広まっている事も一因である。従来のマスメディアは投資効果が直接計測できなかったが、クリック課金型広告は、実際にサイトへ誘導できた人数や購入にいたった人数まで計測できるため、投資効果を時間をおって確認できる。

このような状況を考えると、単に「WWW に延べ何回アクセスがあった」と言う延べ数だけでなく、クッキーなどを使って「WWW に何人からアクセスがあった」とか「その内何人が何回以上アクセスし購入にいたった」と言った異なり数を意識した解析も重要になる。宣伝活動の結果「熱狂的なファンが 2 名からのアクセスが 8 回あった」のと「8 人が 1 回ずつアクセスしてきた」では宣伝の効果(より直接的には売上)が異なる可能性があり、検討を要する。

[3]は、DNS のログデータを利用して映画宣伝の為に解説されたサイトへのアクセス回数を分析し、サイトへのアクセス回数から映画初週の興業業績を推定する手法を報告している。そこでも延べ数より異なり数で計測した値を使った解析の精度が良いと言う結果が得られている。また[3]は DNS を使った計測は読み手の人数を計測している事になり、従来のような恣意的な書き手によるステルスマーケティングの影響を受けない事を利点として述べている

5. おわりに

本発表にあたって研究室所属学生の研究を中心に紹介させていただいた。この記事が少しでも面白いと思っていただけたら、それは彼等の手柄である。面白い研究をしてくれた学生諸氏に感謝したい。

始めに述べたように、異なり数の計測は、概念が簡単である一方、応用範囲が極めて広い。[4]のようにネットワークには全く関係ないような応用にも異なり数の概念自体は使える。今後も新しい応用事例は増えていくものと考える。

[参考文献]

[1] 吉田健一, 三田村健史. "ネットワークデータのオンライン異なり数解析" 人工知能学会誌, Vol.30, No.2, pp.230-237 (2015.3)

[2] Y. Shoumura, Y. Wanatabe, K. Yoshida, "Analyzing the Number of Varieties in Frequently Found Flows" IEICE Transactions on Communication, Vol.E91-B, No.06, pp.1896-1905 (2008)

[3] 三田村健史, 吉田健一, "DNS クエリデータに基づくコンテンツへの関心度分析" 電子情報通信学会論文誌, Vol.J93-B, No.10, pp.1368-1377 (2010)

[4] Kenichi Yoshida, Akito Sakurai, "Short-term Stock Price Analysis Based on Order Book Information" 人工知能学会論文誌, Vol.30, No.5, pp.683-692, (2015.9)