インターネット由来データに対する統計科学的アプローチ

- Web ログデータ・スパムメールの解析 -

南　弘征 (北海道大学 情報基盤センター) / min@cims.hokudai.ac.jp

1　はじめに

インターネット利用者の数、層、利用目的や形態は多岐にわたり、常時、多くの電子情報がネットワーク間で流通している。しかし、見方をかえれば、インターネットは巨大な電子データの生産機構でありライブラリである。たとえば、インターネットアプリケーションとして代表的である World Wide Web のサーバでは、履歴という形で電子データが蓄積されている。また、SPAM メールと俗称される迷惑メールも、見方によっては電子データといえる。このような、インターネットアプリケーションに由来するデータを、本稿では「インターネット由来データ」とよぶ。

インターネット由来データは、その規模や目的から、いわゆるデータマイニング的な技術に依る分析例が多いと思われるが、情報の有用性などに着目すれば、今後、既知のデータ解析手法を駆使した、いわば統計科学的なアプローチが、理論的な裏打ちのある結果として、今後、重要な意味合いを持つと考えることもできる。

本稿では、インターネット由来データに対する統計科学的アプローチとして、Web サーバ履歴に関する解析例、ならびにスパムメールのフィルタリングについて紹介する。

2　Web サーバ履歴の解析

Apache などに代表される Web サーバプログラムは、標準設定でも、アクセス履歴として多くの情報を出力する。インターネット上で売買を行っている企業などでは、顧客管理のために多くの情報をデータベース化し、顧客が再度閲覧した際に適切な内容を表示させるシステムを構築しているが、実現にあたり、閲覧側にも識別のための情報を持たせている。これに対し、本研究室では Web サーバログから単純にどの程度の情報が得られるか、解析を試みている ([2, 3])。ここでは、対応分析による解析例を紹介する。対象としているのは講演者の属する研究室のサーバ (http://tiss1.cims.hokudai.ac.jp/) であり、各利用者別のホームページ空間の他、研究室での歓迎会・送別会などのイベント (event)、研究室近郊を中心とした飲食店に関するページの集合 (eating_out) などをコンテンツとして有している。

表 1 は一定期間のログについて、アクセス元のホスト名から得られたドメイン属性と閲覧されたディレクトリに関する分割表である。このデータに対応分析を適用した結果を図 1 に示す。結果の解釈につ

	eating_out	～min	～mizuta	/	～mazda	～u_ske	event	～hiro	
.ne.jp	11264	393	247	224	267	66	89	48	12598
.or.jp	3213	56	71	83	96	24	16	20	3579
.ac.jp	2602	84	175	157	67	42	21	34	3182
.net	2411	90	75	69	107	8	0	15	2775
.ad.jp	1680	52	42	27	32	0	2	6	1841
.co.jp	1242	108	90	81	69	9	1	0	1600
.jp	427	21	13	9	9	0	0	0	479
.com	235	22	15	21	18	0	0	0	311
計	23074	826	728	671	665	149	129	123	26365

表 1: アクセス元と閲覧先ディレクトリ

いては研究室内で諸説出たが、個人的には第 1 軸が勤勉の度合いを、第 2 軸が一般利用者との関連の強さを示しているのではないかと考える。

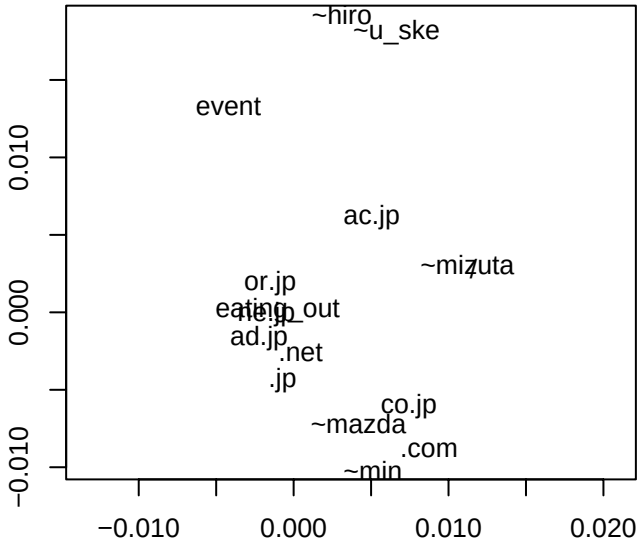

図 1: アクセス元と閲覧先ディレクトリに関する対応分析

この他、表 1 の第 1 列にある、当サーバで圧倒的なアクセス数を誇る `eating_out` の下部ディレクトリ (飲食店メニューのジャンル別、飲食店の住所別) に関する詳細、具体的には一定時間内に同一のアドレスからアクセスしてきたものを「1 利用」として、アクセスしてきた内容についてのデータを作成してクラスタリングを試み、コンテンツに応じた閲覧傾向などについて検討を行っている。なお、直近のデータに関する解析例は当日紹介する。

3　電子メールフィルタリング

俗にスパム (SPAM) と呼ばれる、特定送付先を意図していない広告、誘導などを目的とした電子メールの判別について、[1] などで報告を行っているところである。

3.1　電子メールヘッダーからの情報抽出

研究を始めるにあたり、いわゆる迷惑メール (SPAM) の多くが発信者メールアドレス (`From:`) を偽っている点、宛名 (`To:`) も伏せられていることが多い点などにまず着目した。さらに、配送途中の中継機材がメールに付加していく `Received:`など、一般的なメールソフトウェアでは意図的に調べない限り目に触れないメールヘッダーを精査すると、発信を始めた機材名、所属名などにも偽造の形跡が多く見られた。

電子メールは親書に準じた認識をされていることが多く、メールの内容そのものに極力触れずに SPAM か否かを判断できる方がより望ましいと考えられたため、数百通以上の SPAM を精査し、メールヘッ

ダーに対して見られる傾向を図 2 に示す 22 個のルールとしてまとめ、該当の有無を 0-1 データで表すこととし、判別を試みた結果、最終的にはおよそ 90% 程度の判別率を得ることができた。反面、変数選択の結果、6 変数 (つまり 6 ルール) が除外された (ルール 3,6,16,19,20,22)。ルール設定に問題がなかったかどうか、子細な検討を加え、更なる効率化を検討中である。

1. ある行の Received:フィールド中の 'by' のホスト名と、その上の Received:行の‘ from ’のホスト名が一致しない。
2. 各フィールドに関して、名前 (field-name) はあるが、内容 (field-body) がない。
3. From:や To:などの中身で machine-readable な項が angle brackets(<、>) で囲まれていない。
4. Message-Id:フィールドのドメイン名が無い。
5. ドットで分けられたサブドメインがない。
6. アドレス内にスペース文字がある。
7. (E)SMTP で送っているのに、Received:フィールドに id が無い。
8. Received:フィールドと Received:フィールドの間にそれ以外のヘッダがある。
9. 囲みカッコ（なんでもいい）があるが、中身がない。
10. 宛先が無い。
11. Date:フィールドの時差の項が-1200〜+1200 の範囲から外れた数値になっている。
12. Received:フィールドの‘ from ’の項のホスト名の後に‘ (unverified) ’の文字がある。
13. Message-Id:フィールド自体が無い。
14. Date:フィールドに時差の項が無い。
15. Date:フィールドの時差の項に規則外な文字列がある。
16. Received: フィールド内で、item-name はあるのに中身がない。
17. To: Insured とある以外、宛先がない。
18. Received: フィールドで送信元が名乗ったホスト名が DNS 解決による名前と異なる。
19. Received: フィールドの 'by' の項が具体名であるのに、その上にある Received: 内の from が IP アドレスである。
20. To:, Cc: にあるアドレスが、From:, Sender:, Reply-To: にもある。
21. Received: フィールドの date-time が少ない。
22. Received: フィールドの 'from' の項が構文上不自然である。

図 2: スパムメール判別のためのヘッダー識別ルール

3.2 ベイジアンフィルタの利用と比較

SPAM の除外はインターネット利用者共通の悩みらしく、SPAM メール中に現れる単語の出現確率を蓄積し、各々独自の判定式に代入して得られた指標値が閾値を超えるかどうかで、SPAM の判別を行うようなフィルタリングソフトウェアが最近注目されている。これらは、一応「ベイジアンフィルタ」と総称されているようである。代表的なアルゴリズムは [4] に見られる。

当研究室でも、本文を含んだ解析は行ったが、これらのアプローチとの差異、考え方の違いなどについては当日詳述する。

4　おわりに

本稿で取り上げたほかにも、通信履歴データから不正アクセスを判別するなど、といったさまざまな試みが、ネットワーク由来データに対して、なされているものと承知している[1]。必要性や実用性は十分にあるものと考えられるので、講演者らの属する研究室でも、今後精力的に研究を進めていきたい。

参考文献

[1] Yu. Sato, H. Minami and M.Mizuta (2003): Adaptive Spam Filtering with Text Mining. *Bulletin of the International Statistical Institute, 54th Session, Book 2*, 371–372.

[2] 原田・南・水田 (2004a): Web サーバログ解析によるユーザの特徴抽出およびサイト構造の評価. 2004 年度統計関連学会講演報告集, 14-15.

[3] 原田・小宮・南・水田 (2004b): クラスター分析法による Web サーバログの解析. 日本行動計量学会第 32 回発表論文抄録集, 324-327.

[4] P.Graham: A Plan for Spam. `http://www.paulgraham.com/spam.html`.

[1] 不勉強ゆえ調べきれていないため、このような記載をお許し願いたい。