

ブログ空間における異常訪問行動の分析

山本和紀^{†1} 小野景子^{†2}
熊野雅仁^{†2} 木村昌弘^{†2}

本論文では、ブログ空間における異常行動について調べる。まず、異常行動に関して「アクセス頻度異常」と「趣向比率異常」の2種類の概念を定義する。そして、実際のブログデータを用いて、これらの異常行動を特定し、それらの性質を分析する。特に、趣向比率異常のユーザと人気ブログを発見するイノベータには相関関係があることを示す。

Analyzing anomalous behavior in Blogosphere

KAZUNORI YAMAMOTO,^{†1} KEIKO ONO,^{†2}
MASAHITO KUMANO^{†2} and MASAHIRO KIMURA^{†2}

We investigate anomalous behaviors in Blogosphere. First, we define two concepts for anomalous behavior, the *access frequency anomaly* and the *share anomaly*. Next, using real blog data, we identify these two anomalous behaviors, and analyze their properties. We show in particular that there is a correlation between the share anomaly users and the innovators of finding popular blogs.

1. はじめに

インターネットの普及により、手軽な情報発信ツールとしてブログが普及しており、ユーザのアクセスに着目したブログユーザの行動予測に関する研究が数多く成されている。例えば、アクセスの確率分布と Web ユーザビリティの関係をを用いた研究¹⁾、ブログへのアクセ

ス数と社会現象に関する研究²⁾、ブログのアクセス履歴を可視化により分析する研究³⁾などが報告されている。我々は、アクセスデータを用いて、イノベータ理論に基づいたブログユーザの行動予測の研究⁴⁾、急増する訪問行動の予測の研究⁵⁾を行ってきた。

ところで、推薦システム、e コマース、およびサーバ負荷分散などにおけるコアテクノロジーの一つとして、ブログユーザの大きな行動変化を予測する有効な手法の構築が挙げられる。したがって、本論文では、ブログ空間における異常行動について調べる。まず、異常行動に関して「アクセス頻度異常」と「趣向比率異常」の2種類の概念を定義する。そして、実際のブログデータを用いて、これらの異常行動を特定し、それらの性質を分析する。特に、趣向比率異常のユーザと人気ブログを発見するイノベータの相関関係や、現在において趣向比率異常を行うユーザが将来においても趣向比率異常を行う確率と、現在においてアクセス頻度異常を行うユーザが将来においてもアクセス頻度異常を行う確率との違いなどを調べる。

2. 分析データ

2.1 Doblog データ

本研究では Doblog^{*1} データセット^{*2}のアクセスデータを使用する。アクセスデータは、データ ID、visitor ID、owner ID、アクセス時間で構成されるデータであり、2003 年 10 月から 2005 年 6 月までのデータが存在する。Doblog のデータセットを表 1 に示す。

表 1 Doblog のデータセット
Table 1 Dataset of Doblog

項目	件数
記事数	1,540,077
ユーザ数	53,525
コメント数	2,220,727
トラックバック数	133,177
アクセス数	12,542,581

^{†1} 龍谷大学 大学院 理工学研究科 電子情報学専攻

Division of Electronics and Informatics, Ryukoku University

^{†2} 龍谷大学 理工学部 電子情報学科

Department of Electronics and Informatics, Ryukoku University

^{*1} (株)NTT データ . <http://www.doblog.com/>

^{*2} (株)ホットリンクと(株)NTT データの共同事業契約に基づき、(株)ホットリンクより提供。2003 年 10 月から 2005 年 6 月のデータを利用。

2.2 訪問行動の統計解析

2.2.1 月別アクセス数の推移

図1は、Doblogにおける月別のアクセス数を示したグラフである。徐々にアクセス数が増加し、Doblogが活発になっていることが分かる。

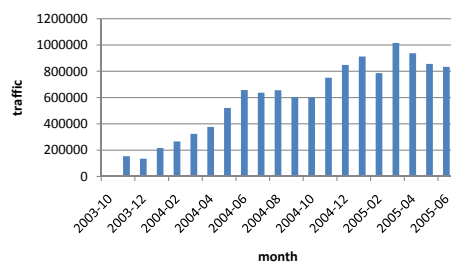


図1 月別アクセス数の推移
Fig.1 Monthly traffic

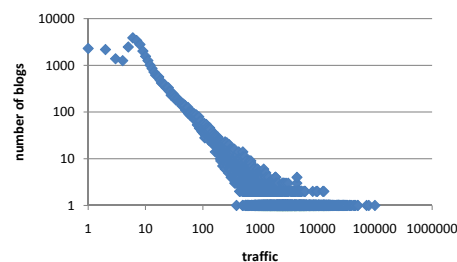


図2 ブログアクセス分布
Fig.2 Distribution of blog access

2.2.2 ブログアクセス分布

図2は、全期間中（2003年10月から2005年6月まで）に、ブログがアクセスされた回数の分布である。一般的なロングテールの分布であることが分かる。

2.3 対象ユーザ

2004年1月1日から2004年12月31日までのアクセスデータをもとに足切りを行い、対象となるユーザを同定した。各ユーザについて、ユーザが任意のブログにアクセスした回数を算出し、平均値以上の値を持つユーザを対象ユーザとした。なお、期間中に任意のブログに1回でもアクセスを行ったユーザは21,628人、平均アクセス回数105.21で足切りをした結果、対象ユーザは2,884人となった。

2.4 対象ブログ

2004年1月1日から2004年12月31日までのアクセスデータをもとに足切りを行い、対象となるブログを同定した。各ブログについて、任意のユーザにアクセスされた回数を算出し、平均値以上の値を持つブログを対象ブログとした。なお、期間中に任意のユーザから1回でもアクセスされたブログは25,193個、アクセスされた回数の平均値95.56で足切りをした結果、対象ブログは3994個となった。

3. 異常訪問行動の分析法の提案

これまでの研究により、ブログへのアクセス数をもとにブログへのアクセスを増加させるユーザの存在について明らかにした。しかしながら、どの程度のアクセス数の増加が起こるのか、その持続性については明らかでない。

図3は、ユーザの1日のブログ訪問回数によって色を変え、プロットしたグラフである。ユーザの1日のブログ訪問回数が少ないほど白、多いほど黒で表現している。図3より、ユーザがブログを訪問する頻度は一様ではないことが分かる。これより、毎日決まったブログにしか訪問しない「通常行動ユーザ」に対し、突然アクセス回数が急増する「異常行動ユーザ」が存在することが分かる。

本研究では、このユーザに着目し、アクセス数の急増を異常行動と捉え定式化を行い、4つの観点から異常行動の分析を行う。

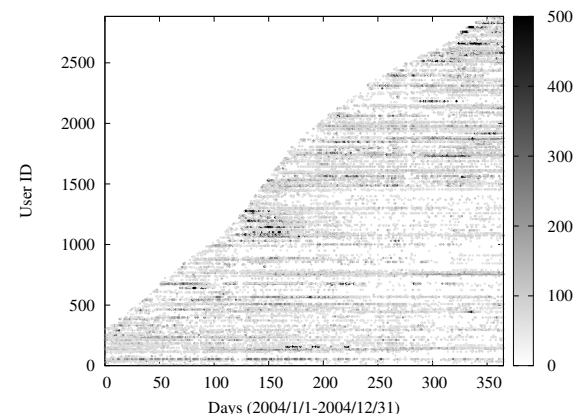


図3 日毎のユーザの行動回数
Fig.3 Daily user's activity

3.1 異常行動の定義

ブログユーザの異常行動を「アクセス頻度」と「趣向比率」の2つの観点から分析を行う。「アクセス頻度」は、ユーザ u_n のブログ b_m へのアクセス回数に着目し、前半期間におけるユーザ u_n のブログ b_m へのアクセス回数 $f_{n,m}(t-1)$ と、後半期間におけるユーザ u_n

のブログ b_m へのアクセス回数 $f_{n,m}(t)$ から，以下の式 1 より異常度 $a_{n,m}(t)$ を算出する．

$$a_{n,m}(t) = \frac{f_{n,m}(t) - f_{n,m}(t-1)}{f_{n,m}(t-1)} \quad (1)$$

異常度 $a_{n,m}(t)$ が 1 を超えている事例を異常な行動事例とする．また，異常な行動を行ったユーザを異常行動ユーザと定義する．「アクセス頻度」における異常行動ユーザは，ブログ b_m への活動量が急増したユーザである．

「趣向比率」については，アクセス回数 $f_{n,m}(t)$ の代わりに，以下の値を用いる．

$$f'_{n,m}(t) = \frac{f_{n,m}(t)}{\sum_m f_{n,m}(t)} \quad (2)$$

$f'_{n,m}(t)$ を「趣向比率」と考え，その値を式 1 に代入し，異常度 $a_{n,m}(t)$ を求める．異常度 $a_{n,m}(t)$ が 1 を超えている事例を異常な行動事例とし，異常な行動を行ったユーザを異常行動ユーザと定義する．「趣向比率」における異常行動ユーザは，訪問するブログの傾向が変化したユーザである．

3.2 異常訪問行動の分析方法

3.2.1 指標 1：行動異常度分布

ある期間における行動の異常度を「アクセス頻度」と「趣向比率」の観点からどの程度存在するのか，分布を示す．

3.2.2 指標 2：異常行動ユーザの恒常性

異常行動ユーザの恒常性を分析を行う．ある期間で異常行動したユーザのうち，次の期間にも異常な行動をするようなユーザ数とその割合を調べことで，アクセス頻度に関する異常行動ユーザと趣向比率に関する異常行動ユーザの恒常性の違いを比較する．

3.2.3 指標 3：ブログの被異常度

ユーザが異常行動したブログについて，ブログ視点で分析を行う．以下の式 3 で，ブログに訪問した異常行動ユーザの異常度の合計をブログ毎に求め，その値をブログの被異常度 A_m とする．

$$A_m(t) = \sum_n a(t) \{a(t); a_{n,m}(t) > 1\} \quad (3)$$

ブログ b_m に関する被異常度 A_m のすべての期間における平均値と分散を求める．

3.2.4 指標 4：ブログに訪問した異常行動ユーザ数に対する全体訪問ユーザ数

ブログ b_m に訪問した異常行動ユーザ数に対して，それ以降の期間に訪問した全ユーザ数

を求める．異常行動ユーザ数と全ユーザ数に相関があれば，異常行動ユーザが相関のあった期間に対し影響を及ぼしていることがいえ，異常行動ユーザに着目することで，将来のブログへのアクセスに関して予測が可能となる．

4. 分析結果

本研究では，図 4 のように，3 か月の期間（period）を 5 つ設定し，前半期間・後半期間の組み合わせ（term）を 4 つとして，分析を行った．

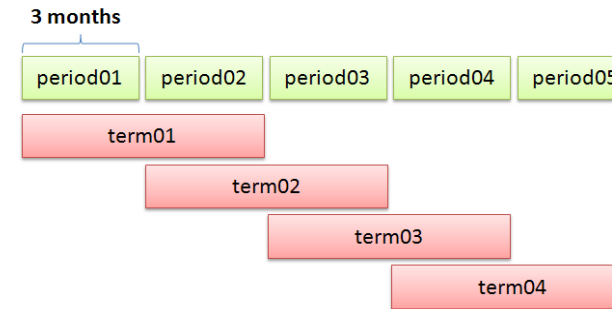


図 4 分析期間
Fig. 4 Period

4.1 指標 1：行動異常度分布

図 5 にアクセス頻度，図 6 に趣向比率について，ともに term02 の期間における異常度の分布を示す．横軸に異常度の下限値，縦軸にアクセス数を示す．横軸 10 における値は異常度が 10 以上の場合のアクセス数を表す．ユーザ u_n がブログ b_m への行動した事例ごとに異常度 $a_{n,m}(t)$ を算出し，異常度 $a_{n,m}(t) > x$ となる事例がいくつ存在するかを示している．

アクセス頻度（図 5）について， $a_{n,m}(t) > 0$ であった事例（アクセスが増加した事例）は 6572 件存在し，そのうち， $a_{n,m}(t) > 1$ であった事例（異常行動事例）は 3055 件存在した．

また，趣向比率（図 6）について， $a_{n,m}(t) > 0$ であった事例（アクセスが増加した事例）が 7280 件存在し，そのうち， $a_{n,m}(t) > 1$ であった事例（異常行動事例）は 2860 件であった．

アクセス頻度，趣向頻度，双方について，異常度 $a_{n,m}(t) > 10$ を超えるような異常行動事例が約 100 件存在していることが分かる．

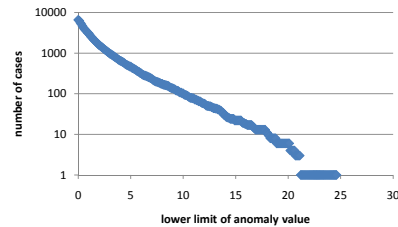


図 5 アクセス頻度に関する，行動異常度分布 (term2)

Fig. 5 Distribution of anomaly action(access frequency)

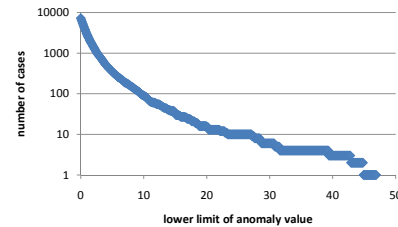


図 6 趣向比率に関する，行動異常度分布 (term2)

Fig. 6 Distribution of anomaly action(share)

4.2 指標 2：異常行動ユーザの恒常性

表 2 は，ある期間において異常行動したユーザが，次の期間以降にも異常行動するユーザ数とその割合を算出した表である．

アクセス頻度，趣向比率ともに，どの期間においても，次の期間にも異常行動するユーザの割合は約 6 割以上であることが分かった．

アクセス頻度において異常行動したユーザよりも，趣向比率において異常行動したユーザのほうが，異常行動を継続する割合が高いといえる．

表 2 異常行動ユーザの恒常性
Table 2 Anomaly user's constancy

	term01	term02	term03	term04
term01(Access frequency)	626(100%)	380(60.7%)	263(42.0%)	200(31.9%)
term01(Share)	664(100%)	478(72.0%)	346(52.1%)	270(40.7%)
term02(Access frequency)	-	1118(100%)	669(59.8%)	482(43.1%)
term02(Share)	-	1291(100%)	892(69.1%)	628(48.6%)
term03(Access frequency)	-	-	1183(100%)	767(64.8%)
term03(Share)	-	-	1517(100%)	1053(69.4%)

4.3 指標 3：ブログの被異常度に関する結果

図 7 はアクセス頻度に関する異常行動ユーザに着目，図 8 は趣向比率に関する異常行動ユーザに着目した時の結果である．ブログ b_m の被異常度 $A_m(t)$ を term01 から term04 の期間で算出し，その平均値が高いブログから順に並べたグラフである．

図 9 はアクセス頻度に関する異常行動ユーザに着目，図 10 は趣向比率に関する異常行動

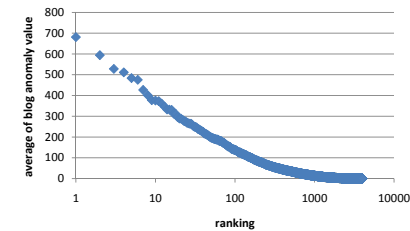


図 7 アクセス頻度に関する，ブログの平均被異常度

Fig. 7 Average of anomaly values(access frequency)

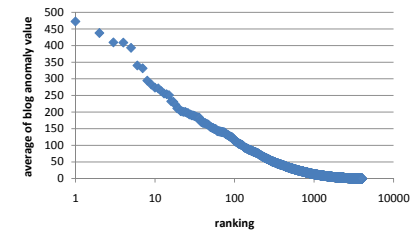


図 8 趣向比率に関する，ブログの平均被異常度

Fig. 8 Average of anomaly values(share)

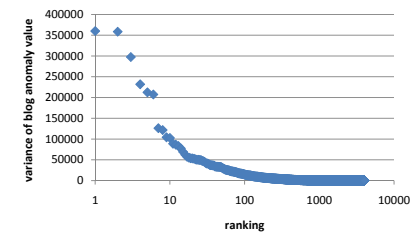


図 9 アクセス頻度に関する，ブログの被異常度の分散

Fig. 9 Variance of anomaly values(access frequency)

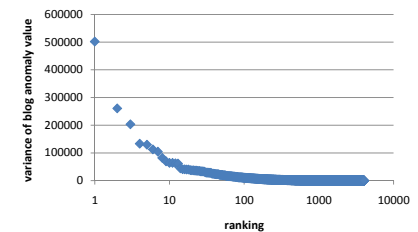


図 10 趣向比率に関する，ブログの被異常度の分散

Fig. 10 Variance of anomaly values(share)

ユーザに着目した時の結果である．ブログ b_m の被異常度 $A_m(t)$ を term01 から term04 の期間で算出し，その分散が大きいブログから順に並べたグラフである．

アクセス頻度，趣向比率の双方において，分散値が高いブログが 10 個程度存在している．これより，期間によって被異常度が急激に変化するようなブログが少ないことがわかった．多くのブログは異常行動され続ける傾向にあることが示唆された．

4.4 指標 4：ブログに訪問した異常行動ユーザ数に対する全体訪問ユーザ数

term1 においてブログに訪問した異常ユーザが他の期間におけるアクセス数に及ぼす影響を考察する．

図 11，図 12 は，term01 においてブログ b_m に訪問した異常行動ユーザ数と，period01 においてブログ b_m に訪問した全ユーザ数との関係を表している．図 11 はアクセス頻度，図 12 は趣向比率に関するグラフであるが，どちらも相関がない．

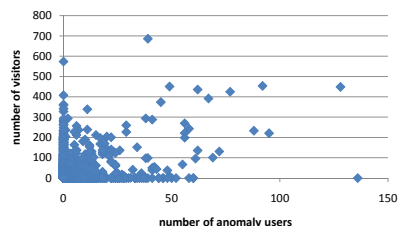


図 11 アクセス頻度に関する，ブログに訪問した異常行動ユーザ数に対する全体訪問ユーザ数 (period01)

Fig.11 Relation between the number of anomaly users and the number of visitors in blog(access frequency)(period01)

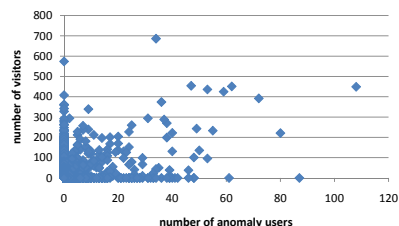


図 12 趣向比率に関する，ブログに訪問した異常行動ユーザ数に対する全体訪問ユーザ数 (period01)

Fig.12 Relation between the number of anomaly users and the number of visitors in blog(share)(period01)

図 13, 図 14 は, term01 においてブログ b_m に訪問した異常行動ユーザ数と, period02 においてブログ b_m に訪問した全ユーザ数との関係を表している. term01 においてブログ b_m に訪問した異常行動ユーザ数と, period02 においてブログ b_m に訪問するユーザ数には相関がある.

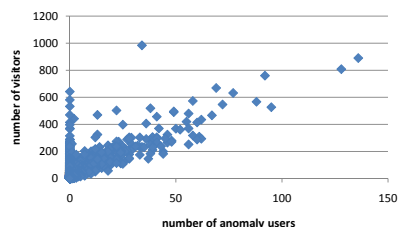


図 13 アクセス頻度に関する結果 (period02)
Fig.13 Access frequency(period02)

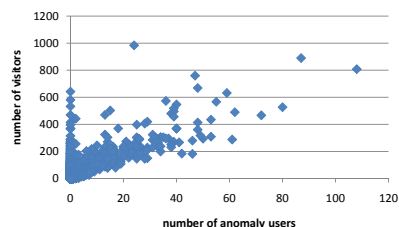


図 14 趣向比率に関する結果 (period02)
Fig.14 Share(period02)

図 15, 図 16, 図 17, 図 18 は, term01 においてブログ b_m に訪問した異常行動ユーザ数と, period03 または period4 においてブログ b_m に訪問した全ユーザ数との関係を表している. term01 においてブログ b_m に訪問した異常行動ユーザ数と, period03 においてブログ b_m に訪問するユーザ数には中程度の相関があり, period4 では相関がないことが分かる. 異常行動ユーザ数とブログに訪問するユーザ数には相関関係があり, 相関は期間によって

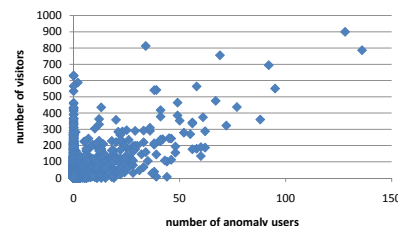


図 15 アクセス頻度に関する結果 (period03)
Fig.15 Access frequency(period03)

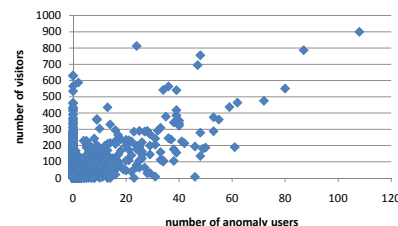


図 16 趣向比率に関する結果 (period03)
Fig.16 Share(period03)

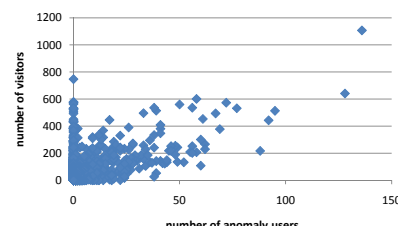


図 17 アクセス頻度に関する結果 (period04)
Fig.17 Access frequency(period04)

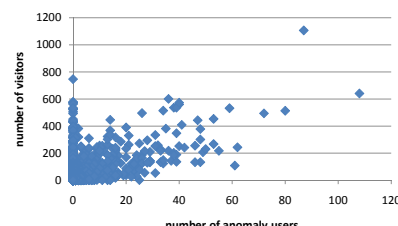


図 18 趣向比率に関する結果 (period04)
Fig.18 Share(period04)

異なることが分かった. term01 においてブログ b_m に訪問した異常行動ユーザ数に対して最も相関があるのは, period02 におけるブログ b_m に訪問した全ユーザ数である. それ以降の期間になると相関が薄れていく傾向が見られた. 異常行動ユーザに着目することで, 直後の訪問ユーザ数を予測できるといえる.

異常行動ユーザと直後の訪問ユーザ数の相関が高いことが分かった. ブログにおいて異常行動ユーザが訪問を始める時間との関係性について調べた.

図 19 はアクセス頻度に関する異常行動ユーザが訪問し始めた時間を, 図 20 は趣向比率に関する異常行動ユーザを, それぞれ, ブログ b_k の日毎のアクセス数とともに示した. ブログ b_k は, 2004 年 6 月頃から急激にアクセス数が伸びている. これらの結果より, イノベータ理論におけるイノベータ的な特徴が現れていることがわかる. 異常ユーザのユーザ数と訪問するユーザ数との間に見られた相関関係は異常ユーザがイノベータ的な役割を果たすことに起因していると考えられる.

アクセス頻度に関する異常行動ユーザ (図 19) は 2004 年 4 月頃から訪問し始めている

のに対し、趣向比率に関する異常行動ユーザ（図 20）は 2004 年 2 月頃から訪問し始めていた。これより、アクセス頻度に関する異常行動ユーザよりも、趣向頻度に関する異常行動ユーザのほうが、よりイノベータ的な要素を持っていることが分かった。

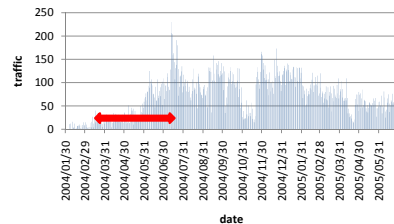


図 19 あるブログにおける日毎のアクセス数と、アクセス頻度に関する異常行動ユーザの新規訪問

Fig.19 Relation between the number of accesses and the first day of the anomaly user access(Access frequency)

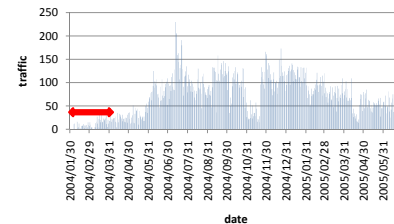


図 20 あるブログにおける日毎のアクセス数と、趣向比率に関する異常行動ユーザの新規訪問

Fig.20 Relation between the number of accesses and the first day of the anomaly user access(Share)

5. ま と め

本研究では、ブログ空間における異常訪問行動の分析を行った。まず、異常行動に関して、「アクセス頻度異常」と「趣向比率異常」の 2 種類の概念を定義した。「アクセス頻度」と「趣向比率」という 2 つの視点から、日毎のユーザの行動回数と行動異常度分布を求めることにより、アクセス頻度異常行動と趣向比率異常行動が多数存在することを観測した。また、趣向比率異常のユーザと人気ブログを発見するイノベータの間に相関関係があることや、現在において趣向比率異常を行うユーザが将来においても趣向比率異常を行う確率のほう、現在においてアクセス頻度異常を行うユーザが将来においてもアクセス頻度異常を行う確率よりも高いことを観測した。

さらに、ブログの被異常度を算出し、異常行動ユーザがどのブログに異常行動をしたかを分析した。ブログの被異常度の分散において、期間による分散値が大きいブログは少なく、分散値が小さいブログがほとんどであったことから、ブログの被異常度は時期によってあまり変化しないことが観測された。すなわち、異常行動されやすいブログは継続して異常行動され、異常行動されにくいブログは常に異常行動されにくいことが示唆された。

参 考 文 献

- 1) 山縣 修, 柳下 孝義: アクセス確率分布と Web ユーザビリティの関連の評価, The bulletin of Health Science University (5), pp.33-44, (2009) .
- 2) Heather A. Johnsona, Michael M. Wagnera: Analysis of Web Access Logs for Surveillance of Influenza, MEDINFO 2004 Amsterdam: IOS Press, pp.1202-1206(2004) .
- 3) Takayuki Itoh, Member, IEEE Computer Society, Yumi Yamaguchi, Yuko Ikehata, and Yasumasa Kajinaga : Hierarchical Data Visualization Using a Fast Rectangle-Packing Algorithm, IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 10, NO. 3, pp.302-313(2004) .
- 4) 山本 和紀, 伊藤 政志, 熊野 雅仁, 木村 昌弘: イノベータ理論を用いたブログユーザの行動予測, ネットワーク生態学シンポジウム (NetecoSymp 2009), P9-10 (2009) .
- 5) 山本 和紀, 熊野 雅仁, 木村 昌弘: アクセス履歴を用いたブログ空間における急増する訪問行動の予測, 情報処理学会 50 周年記念全国大会講演論文集, 5A-3 (2010) .
- 6) 総務省情報通信政策研究所 (IICP) 調査研究部: ブログの実態に関する調査研究の結果 (2008) .