DNSクエリデータに基づくコンテンツへの関心度分析

三田村健史[†,††a)]　吉田　健一[††]

DNS Query Analysis for Sociology

Takeshi MITAMURA[†,††a)] and Kenichi YOSHIDA[††]

あらまし　本論文では，DNSクエリデータを用いたコンテンツに対する社会からの関心度の測定手法を提案する．従来の手法は，主にインターネット上の仮想コミュニティへの投稿者という「書き手側」の情報を分析することで関心度の測定を行うが，そこには「書き手側」の意図的な情報操作による影響を受けやすいという課題があった．そこで本論文では，意図的な情報操作の難しいDNSクエリデータを用いた「読み手側」からの関心度測定手法を提案する．提案手法では，DNSキャッシュの影響を排除するためDNSクエリデータの異なり数を用いる．また，提案手法を検証するために日本映画の公開初週の観客動員数予測を行い，公開前2か月間における予測精度の推移についても考察し，提案手法が実務へ適用できる精度をもつことを示す．

キーワード　DNS，関心度，DNSキャッシュ，異なり数，観客動員数予測

1. まえがき

近年，インターネット上のコミュニティ形成を支援するブログやSNSなどのサービスが，日常的に利用されるようになってきた．このようなインターネット上のコミュニティ（以下，仮想コミュニティ）での現象と現実社会での現象との関係についての研究は多く行われている．

例えば[1]は，書籍に言及しているブログへの書込み数の時系列推移が，当該書籍の電子商取引サイトでの売り上げランキングへ及ぼす影響は1.7日から8.8日後に現れるとしている．また[2]は，仮想コミュニティとしてインターネット掲示板を用い，そこへの書込みからテレビCMに関する語彙群を抽出し，その時系列データから現実社会でのCMにおける関心度を議論している．

これらの研究に代表されるように，仮想コミュニティにおける現象と現実社会での現象の関連性に関する研究は，主として仮想コミュニティにおける「書き手側」の視点から記述内容や記述頻度，時間などの特徴を抽出し，現実社会での現象との関係を分析することが主流である．

ここでの疑問は，「仮想コミュニティにおける書き手は，作為的な書き手ではないといえるのだろうか？」ということである．例えば，インターネット掲示板への書き手が，CMに関する好意的な書込みを掲示板に連続投稿し続けることで関心度を見かけ上長引かせ，CMによる効果が継続しているように見せかける操作を行うことは，技術的には可能である．特に，分析する内容（商材や人物など），分析する掲示板やブログなど分析対象などを絞り込むほど，「書き手側」による作為的な影響は大きくなると考えられる．

「書き手側」による作為的な影響事例としてスパムブログがある．スパムブログは，大量の書込みにより悪意ある特定サイトに誘導することなどをねらったブログのことである．例えば[3]は，20日間の約1500万ブログ更新データを調査した結果，約75%がスパムブログによる更新データであり，それは全英語ブログサイトの約88%を占めるとしている．また[4]は，スパムブログによる悪影響として(a)情報検索品質の低下，(b)ネットワークとストレージの浪費，を示している．

本研究では，このようなコンテンツへの「書き手側」

† (株) 日本レジストリサービス，東京都
Japan Registry Services Co., Ltd., Chiyoda First Bldg. East 13F, 3–8–1 Nishi-Kanda Chiyoda-ku, Tokyo, 101–0065 Japan
†† 筑波大学大学院ビジネス科学研究科，東京都
Graduate School of Business Sciences, University of Tsukuba, 3–29–1 Otsuka, Bunkyo-ku, Tokyo, 112–0012 Japan
a) E-mail: mitamura@jprs.co.jp

を中心とした関心度分析法のもつ課題を解決する手法を提案する．具体的には，インターネットにおける現象と現実社会での現象の関係において，作為的な意思が入りにくい「読み手側」を中心とした関心度分析法を提案する．提案する手法では，「読み手側」の関心を表す情報としてDNSへのクエリデータを用いる．

一般的にDNSへのクエリは，ウェブの閲覧やメールの送受信などで発生する．つまり，コンテンツの読み手がアクセスする場合にDNSへのクエリは常に発生する．またDNSへのクエリは，特定のブログや掲示板などにアクセスする場合にのみ発生するものではなく，すべてのウェブのアクセスで共通に発生するものである．このことから，DNSへのクエリから抽出したデータには，コンテンツへの「書き手側」の意思は入りにくいといえる．

また，本研究では提案手法検証のための実例として，日本映画への関心度測定を取り上げ，次の二つを研究目的とする．

- 「書き手側」ではなく，「読み手側」のコンテンツへの関心度を測定する手法を明らかにする．特に，その過程においてDNSというインターネットの技術が，現実社会の分析に利用できることを検証する．
- 実事例として日本映画を取り上げ，その公開初週の観客動員数の予測を行うことで，提案手法により関心度の測定が可能であることを示す．

本論文の構成は次のとおりである．まず**2.**で関連研究を概観し本研究の位置付けを明確にする．**3.**で提案手法を説明し，次いで**4.**にて提案手法による分析結果の例を報告する．**5.**では提案手法の実務への適用について考察し，最後に**6.**で結論を述べる．

2. 関連研究

現実社会での現象と仮想コミュニティにおける「書き手側」の現象との関係分析は，主に両者で共通となる基軸（例えば時間や人，出来事など）を合わせ，その基軸上での現実社会の現象と仮想コミュニティにおける「書き手側」の書込み内容との関係を特徴づけるという手法を用いたものが主流である．

例えば，現実社会の現象として消費者行動に注目し仮想コミュニティとの関係を分析した研究として，実際の消費者の購買行動とブログや掲示板に対する書込み内容の関係分析を行う研究が盛んである．仮想コミュニティのユーザと非ユーザを比較しオンライン・ショッピングサイトの利用率やリピート率の違いを示した研究[5]や，仮想コミュニティの存在と消費者の購買行動におけるブランド力の関係を分析した研究[6]，株式市場の収益率や取扱高の時系列の推移と個人投資家の掲示板への書込み頻度との相関を分析した研究[7]などである．

消費者行動に結び付く意思決定と仮想コミュニティの関係を情報伝搬の関係から研究するものも重要である．例えば[8]は，インターネット上の評価情報に関して情報を発信する者(Radical Access Member)と黙って読んでいる者(Read Only Member)に分類し，各々の購買行動に及ぼす影響を明らかにしている．

これら以外にも仮想コミュニティの書込みに着目し，そこから現実社会での現象に対する関心度を抽出しようとする研究事例は多い．例えば，ブログ間の関係においてトピックがどのように伝達されていくかをモデル化した研究[9]や，複数の国のWeblog記事や新聞記事，メールマガジンなどからトピックを単語レベルで抽出し，その時間軸上の推移と気温など現実社会での出来事の推移との相関を見る研究[10]，2ch上の視聴者コミュニティで展開される対話文をもとにドラマ番組に対する視聴者の注目状態の特徴パターンを抽出する研究[11]，ブログから街の話題に関する言葉を集め，その街における人の体験として集約し地図上にマッピングするという関心度と経験の対応付けの研究[12]などがある．

これら現実社会での現象と仮想コミュニティでの現象の関係を扱った研究の多くは，両者の間に強い相関があることを示している．しかし，今までの研究は主に仮想コミュニティにおける「書き手側」の情報分析にとどまっており，「書き手側」の意図的な情報操作による影響を受けやすいという課題を抱えている．このような背景から本論文では意図的な情報操作の影響を受けにくい「読み手側」からの関心度の測定手法を提案する．

本研究では，提案手法の有効性を検証するために日本映画への関心度分析を試みた．分析対象である日本映画は，2007年に閣議決定された「文化芸術の振興に関する基本的な方針（第2次基本方針）(注1)」や「国立メディア芸術総合センター（仮称）構想(注2)」で取り組むべき重要コンテンツの一つとして取り上げられ

(注1)：http://www.bunka.go.jp/bunka_gyousei/housin/pdf/kihon_housin_2ji.pdf
(注2)：http://www.bunka.go.jp/oshirase_other/2009/mediageijutsu_090514.html

ており，近年その関心度分析の重要性が高まっている．

具体的な関心度のメジャーとして選択した映画の興行成績を予測する研究は，主に米国においてマーケティング的な観点で実施されている．例えば，当初の潜在顧客規模 (N_0)，そのうち実際に観客となる割合 (P_0)，公開前の口コミ効果 (σ) という三つの変数から興行収入のモデル化を試みる研究 [13]，インターネット掲示板への発言量と評論家のレビュー数が映画興行の初期の興行成績に対して与える影響の研究 [14]，映画の興行収入の時系列推移において，映画ジャンルや公開時期などの変数に加え，ブログでの映画に対する評価コメントが興行収入に及ぼす影響の研究 [15] などがある．

これらの研究は，主に多くの情報の中から興行成績に，より影響を与える説明変数を探し出す内容となっている．例えば [15] は，ブログへの書込み内容と興行収益の相関が，公開前で 0.454～0.542，公開後で 0.478～0.614 程度であることを示している．本研究では DNS クエリデータを用い，これより高い精度での予測が可能であることを示す．

3. 異なり数を用いた関心度測定手法の提案

本章では，読み手側からの関心度測定手法として，DNS クエリデータの異なり数を用いた関心度測定手法を提案する．

3.1 システム構成

図 1 に，一般的な DNS クエリの仕組みと，本研究における関心度測定システムの構成を示す．この事例では，まずウェブの閲覧者によるブラウザの操作により，ISP の DNS サーバにウェブの URL である www.example.jp と IP アドレスの変換が要求される．これに対し ISP の DNS サーバでは，m.root-servers.net から．jp のデータをもつ JP-DNS サーバ（例えば a.dns.jp）の IP アドレスを取得し，次に a.dns.jp から example.jp のデータをもつ dns.example.ad.jp の IP アドレスを取得する．最後に dns.example.ad.jp から www.example.jp の IP アドレスを取得し，最終的にブラウザが動作する PC に www.example.jp の IP アドレスを通知する．

本研究における分析データの取得は，統計情報を公開している代表的な JP-DNS サーバ（図 1 の a.dns.jp/g.dns.jp）で行った．a.dns.jp は，IPv4/IPv6 アドレスで IP anycast 技術を用い東京/大阪の複数

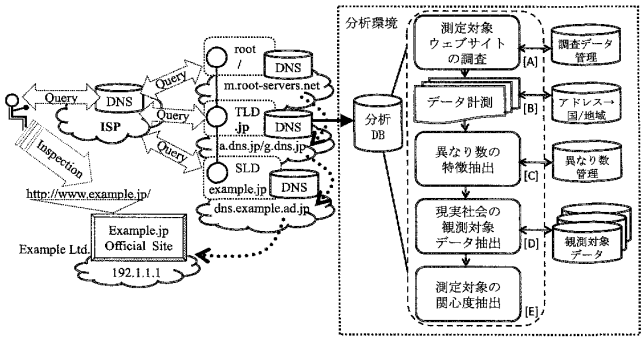

図 1 関心度測定システム
Fig. 1 Degrees of concern measurement system.

拠点で運用(注3)されている．g.dns.jp は，IPv4 アドレスで unicast 技術を用い東京拠点で運用されている．a.dns.jp，g.dns.jp の各々で取得されたログデータは，一つのログデータとして統合した．統合したログデータは，1 日 1 回のバッチ処理で，関心度測定のための分析環境に転送し，分析用のデータベースに蓄積した後，関心度分析に使用した．具体的には，図 1 に示す分析環境において，関心度測定のためのデータ処理を次の流れで実施した．

① 関心度測定を行う対象との関係が一意に特定できるウェブサイト（例えば，商品専用サイト，個人サイト，組織・団体サイトなど）を特定し，調査データ管理データベースに登録する．(図 1 中 [A])

② a.dns.jp/g.dns.jp から転送されたログデータを格納するデータベース（図 1 の分析 DB）に対し，調査データ管理データベースに登録されたウェブサイトのドメイン名について検索を行う．検索結果から参照元のアドレス情報（ドメイン名，IP アドレス，AS 番号など）を取り出し，そのアドレスが登録/管理された国名/地域名との対応データベースを検索し異なり数（次節を参照）を計測する．(図 1 中 [B])

③ 整理し特徴を抽出する．(図 1 中 [C])

④ ウェブサイトとかかわりがある現実社会における現象の観測データ（例えば，商品の売上高，個人や組織・団体などの活動記録など）を調査し，観測対象データベースに格納する．(図 1 中 [D])

⑤ ③と④の結果を使って，分析時間軸を合わせて回帰分析を行い関心度の算出式を立てる．(図 1 中 [E])

3.2 延べ数と異なり数

本研究では，コンテンツへの関心度を測定するため

(注3)：http://jprs.jp/tech/jp-dns-info/2008-10-06-jp-dns-servers.html

の DNS クエリデータとして，図 2 に示す 2 種類の変数を用いる．一方は延べ数であり，もう一方は異なり数である．延べ数はウェブサイトへの DNS クエリが何回発生したかという累計を示す数であり，異なり数はウェブサイトへの DNS クエリが何箇所から発生したかという種類を示す数である．更に異なり数では，IP アドレスと国・地域の 2 種類の変数を用いる．前者は何台の機器から DNS クエリが発生したかを示し，後者は何か国から発生したかを示す．

3.3 DNS キャッシュの影響と異なり数

DNS はキャッシュという機能をもっている．これは，ドメイン名単位に決められた TTL(注4) という時間の間，検索されたドメイン名と IP アドレスの対応データを保持し，TTL 満了までの間に同じドメイン名に対するクエリを再度受けた場合に保持している情報を代理応答する機能である．この機能はキャッシュサーバで提供され，図 1 の例では ISP の DNS サーバで実行される．**3.1** で示したように，本提案手法におけるデータ取得は JP-DNS サーバで行い，そこへのクエリは ISP の DNS サーバなどキャッシュサーバから送信される．そのため，ブラウザからの DNS クエリに対し，キャッシュサーバによる代理応答が行われる可能性があり，JP-DNS サーバでは必ずしもすべてのクエリを受信できているとはいえない．つまり，JP-DNS サーバにおける単純なクエリの延べ数計測では，キャッシュによる影響を受け，正しく関心度を測れない可能性がある．

提案手法では，何回アクセスされたかという延べ数より，何箇所からアクセスされたかを表す異なり数の方が，DNS キャッシュの影響を受けにくいと考え，単純な DNS クエリの延べ数ではなく異なり数を用いて関心度を測定する．

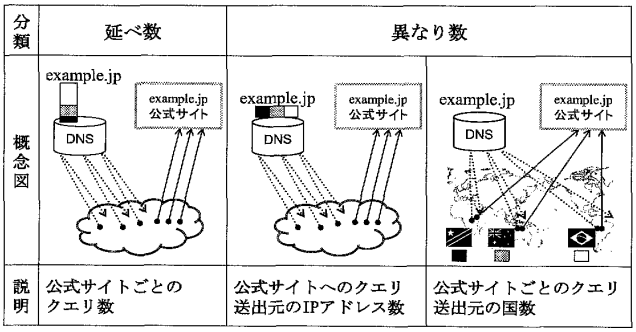

図 2　変数の定義
Fig. 2　Definition of variable.

4. 提案手法を用いた分析結果

前章では，関心度測定に DNS クエリデータの異なり数を用いる手法を提案した．本章では提案手法を用いた事例検証を行う．検証で取り上げるコンテンツとその関心度は，「日本映画とその観客動員数」である．

4.1 日本映画の観客動員数予測分析

表 1 に観客動員数予測分析対象を示す．関心度測定において，映画公開初週とした理由は次の二つである．

(a)　映画の公開初週の興行収入（観客動員数）を決定できれば，高い精度で全興行収入を予測できる [16] ことや，全興行収入の 25%は公開初週の 2 週間の収入が占める [17] など，興行成績における公開初週の重要性が示されているため．

(b)　映画公開前では主な情報源は映画の公式サイトとなるが，映画公開後になるとその情報源は広がることが推測され，観客動員数に影響を与える要素が多くなりノイズが増加すると考えられるため．

4.2 分析の流れとデータ抽出範囲

提案手法による分析を以下の流れで行った．

分析 1：単独変数による説明力分析

観客動員数とクエリ数，IP アドレス数，国数の 3 変数を用いて単回帰分析を行う．ここでは，延べ数，異なり数の説明力と DNS キャッシュによる影響について分析を行う．

分析 2：変数の組合せによる重回帰分析

DNS キャッシュの影響を考慮し，IP アドレス数，国数の組合せで重回帰分析を行う．

分析 3：回帰式の説明力の改善

分析 2 で実施した重回帰分析結果において，実務へ

表 1　分析対象
Table 1　Analysis object.

映画	
作品	.jp ドメイン名の公式サイトを持つ日本映画
公開時期	2008 年 6 月～2009 年 11 月
作品数	50 作品
観客動員数	公開初週の動員数
DNS のクエリデータ	
分析期間	2007 年 7 月～2009 年 11 月
データ取得	JP-DNS のクエリ（a.dns.jp，g.dns.jp）

(注4)：DNS には各ドメインに対する運用方法を記述したゾーンファイルという情報が格納されており，そこで TTL（Time To Live）を指定する．

の適用を考慮して事例固有の説明力改善を行う．

本分析におけるデータ抽出範囲を表2に示す．

4.3 計測データの説明力分析

まず，クエリ数，IPアドレス数，国数の各々の説明力を分析するために，各変数を用いた単回帰分析を行った．単回帰分析の結果を表3に示す．

相関係数はIPアドレス数＞国数＞クエリ数となった．このことから，延べ数（クエリ数）よりも異なり数（IPアドレス数と国数）の方が高い説明力を示す傾向があることが分かった．

また，表4に示すように，クエリ数とIPアドレス数は相関係数が高く，重回帰分析時に多重共線性を起こす可能性がある．これらの結果から，次のことが明らかとなった．

(a) 映画の公開初週の観客動員数は，異なり数と正の相関がある．

(b) 異なり数は延べ数よりも説明力が高い傾向にある．

(c) IPアドレス数とクエリ数は，相関係数が高い傾向にある．

表2 抽出範囲
Table 2 The range of extraction.

クエリ数	クエリ初観測日から映画公開前日までに発生したクエリ累計数
IPアドレス数	クエリ初観測日から映画公開前日までの各日にアクセスしてきたIPアドレスの種類
国数	クエリ初観測日から映画公開前日までの各日にアクセスしてきた国の種類

表3 単回帰分析結果
Table 3 Simple linear regression result.

	延べ数	異なり数	
	クエリ数	IPアドレス数	国数
相関係数	0.644	0.776	0.695
標準誤差	139843.3	115335.9	131508.6
平均	87607.4	45320.0	83.3

表4 変数間の相関係数
Table 4 Correlation coefficient between variables.

	クエリ数	IPアドレス数	国数
クエリ数	1	−	−
IPアドレス数	0.763	1	−
国数	0.570	0.535	1

4.4 重回帰分析の結果

4.3 の結果を踏まえ，説明変数に国数とIPアドレス数（いずれも異なり数）を用いた重回帰分析を行った．結果を図3，図4，表5に示す(注5)．

まず「国数+IPアドレス数」の相関係数と自由度調整済み決定係数であるが，図3，図4に示すように「クエリ数」「IPアドレス数」「国数」の単回帰結果よりもともに大きく改善している．また表5に示すように，説明変数である国数とIPアドレス数の回帰式における各々の影響力は，IPアドレス数（t値：6.106）の方が国数（t値：4.222）よりも大きいという結果になった．

予測式は次のとおりである．

$$\mathcal{Y} = 3457.257\mathcal{X}_1 + 3.807515\mathcal{X}_2 - 256791$$

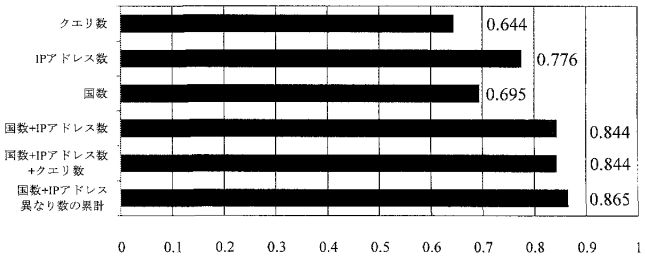

図3 相関係数比較
Fig. 3 Correlation coefficient comparison.

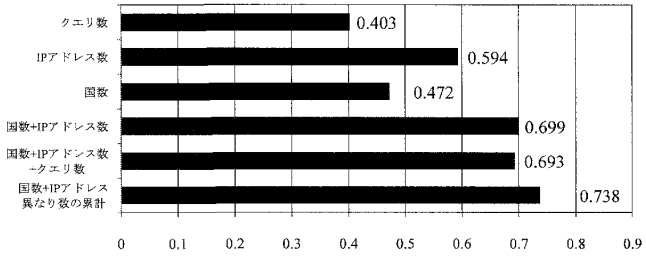

図4 自由度調整済み決定係数比較
Fig. 4 Adjusted R^2 comparison.

表5 回帰係数とt値
Table 5 Regression coefficient and t value.

		国数+IPアドレス数		国数+IPアドレス異なり数の累計	
		回帰係数	t値	回帰係数	t値
説明変数	切片	-256791	-4.231	-248298	-4.438
	国数	3457.257	4.222	3120.225	4.031
	IPアドレス数	3.808	6.106	−	−
	IPアドレス異なり数の累計	−	−	5.298	7.053

(注5)：「IPアドレス異なり数の累計」については **4.5** を参照．

表 6　回帰係数と t 値（変数：延べ数追加）
Table 6　Regression coefficient and t value.　(variables: total number addition)

	回帰係数	t 値
切片	-259047	-4.232
国数	3517.432	4.061
IP アドレス数	3.937	4.695
クエリ数	-0.099	-0.234

ここで，

$\mathcal{Y}$：日本映画の公開初週の観客動員数

$\mathcal{X}_1$：国数（異なり数：公開前日時点）

$\mathcal{X}_2$ ：IP アドレス数（異なり数：公開前日時点）

参考までに「国数+IP アドレス数」に「クエリ数」を変数として加えた場合の回帰分析結果も図 3，図 4，表 6 に示す．図 3 に示すように相関係数は同程度（両者とも 0.844）の数値となっているが，自由度調整済み決定係数の数値は図 4 に示すように 0.699 から 0.693 へと下がっており説明力は落ちている．また表 6 に示すようにクエリ数の t 値は小さく多重共線性の疑いが強い．これは関心度測定において，延べ数よりも異なり数を用いた提案手法の有用性を示している．

4.5　予測精度改善の試み

本節では実務への適用を考慮し，興行予測を行う場合における予測精度の改善を試みる．改善は，単回帰分析で説明力の最も高かった変数である IP アドレス数に着目する．

例えば図 5 のグラフは，ある映画の IP アドレス数の時系列データである．縦軸は IP アドレス数，横軸は DNS クエリの初観測日から映画公開前日までの日数である．この時系列データの推移を見ると，公開日に近づくにつれて IP アドレス数は増加しているが，傾向には段階的な特徴が見られる．公開日 5 か月以上前ではほとんどデータ観測が見られず，5 か月前から 3 か月前の間でいったん緩やかな増加となり，2 か月前から 1 か月前で急激な増加傾向に変わる．

これらの傾向を踏まえ本改善の試みでは，IP アドレス数のデータ抽出において，「公式サイトに対し，より映画公開日に近く，より多くの国からアクセスがある日は，関心度に関する分析に有効な情報が，より集まっている可能性がある」という仮説を置いた．この仮説に基づき，「より映画公開日に近く，国数が平均よりも多い日が連続している期間」の IP アドレス異なり数の累計を計測することで予測精度の改善を試みた．

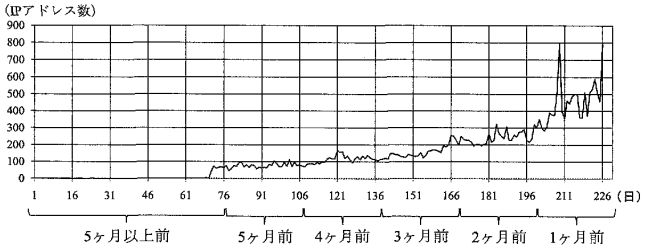

図 5　IP アドレス数の時系列データ例
Fig. 5　The longitudinal data example of IP address number.

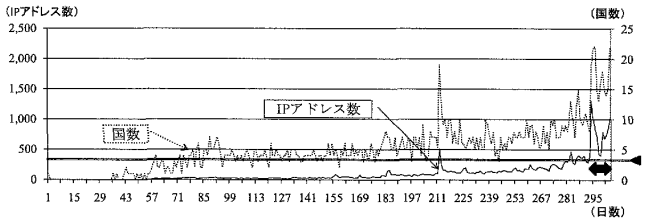

図 6　データ抽出領域
Fig. 6　Data extraction area.

IP アドレス異なり数の累計の抽出領域を図示した事例が図 6 になる．図 6 の左縦軸は IP アドレス数，右縦軸は国数，横軸は初クエリ観測からの日数である．右縦軸 4.8 の横線が国数の平均値を示し，それを超える公開日により近い連続した期間の IP アドレス異なり数の累計（図中右端の両矢印で示した期間）が，改善を試みるデータ抽出対象領域である．

このような方針を踏まえ IP アドレス異なり数のデータ抽出を全分析対象作品に対して行い，「国数+IP アドレス異なり数の累計」での回帰分析を行った．この結果も図 3，図 4，表 5 に示す．

まず相関係数と自由度調整済み決定係数の変化は，図 3，図 4 に示すようにそれぞれ 0.844 から 0.865 へ，0.699 から 0.738 へとともに約 3%改善した．また改善された回帰式において表 5 に示す t 値を見ると，「IP アドレス異なり数の累計」の t 値が「IP アドレス数」の t 値よりも大きく伸びており，仮説は妥当であったといえる．

改善を行った場合の予測式は次のとおりである．

$$\mathcal{Y} = 3120.225\mathcal{X}_1 + 5.297912\mathcal{X}_2 - 248298$$

ここで，

$\mathcal{Y}$：日本映画の公開初週の観客動員数

$\mathcal{X}_1$：国数（異なり数：公開前日時点）

$\mathcal{X}_2$：IP アドレス異なり数の累計

（ただし，国数が平均値を超えた連続期間の累計）

5. 考　　察

4. では，提案手法を用いて映画公開初週の観客動員数の予測を行い，提案手法の有効性を検証した．従来手法 [15] の予測精度が相関係数で 0.478～0.614 であるのに対して，提案手法では 0.865 と大きく改善している．

本章では，**4.** で示した日本映画の観客動員数予測式の実務への適用可能性について考察する．

5.1 実務への適用可能性

本研究の実務適用可能性を探るために，実務家へのインタビューを行った．映画配給会社によっても差異は出ると考えられるが，公開前の主な決定事項と時期は次のような目安であるとのことであった．

- 公開 2 週間程度前まで：シネマコンプレックスの席数決定
- 公開 1 か月程度前まで：メディアへの広告枠決定
- 公開 3 か月程度前まで：上映館（施設）の決定

このことから，実務では最低でも映画公開の 2 週間前，可能であれば 1 か月から 3 か月くらい前の時点での観客動員数の予測を求められることが分かった．

実務への適用可能性を検証するため，データ計測時期を映画公開前日から 2 週間単位で 2 か月前までさかのぼり，**4.5** で示した予測式の精度がどのように推移するのかを分析した．今回分析対象とした映画作品は，初クエリの発生から映画公開前までの期間において 3 か月に満たない作品があったため，分析は 2 か月前までとした．自由度調整済み決定係数の推移を図 7 に，観客動員数予測の推移を図 8 に示す．

まず，自由度調整済み決定係数の推移であるが，図 7 に示すように映画公開の 1 か月前で 0.608 と高い説明力を示す結果となった．

観客動員数予測の推移については，図 8 に示す予測と実測が一致する実線を軸に分布を見ると，0.5 か月前の段階で観客動員数の大きな作品から外れ始め，映画公開前日からさかのぼるほど，実測値よりも予測値が低くなる傾向を示す結果となった．

図 7，図 8 から全体的な推移傾向は，公開前日から 1 か月前までは緩やかに下がり，1 か月前と 1.5 か月前の間で落込みが大きくなり，その後再度緩やかに下がることが分かった．

次に，もう少し詳細に観客動員数の実測と予測の差異について分析する．分析対象とした映画の観客動員数に関する基本統計量を表 7 に示す．

この基本統計量を踏まえ，観客動員数の標準誤差である 2.5 万人単位，及び平均値と標準誤差の比率である 12.5 %単位の差異を基本として A から F までの六つに分類した．

A：予測と実測の差異が 2.5 万人未満，または 12.5%未満

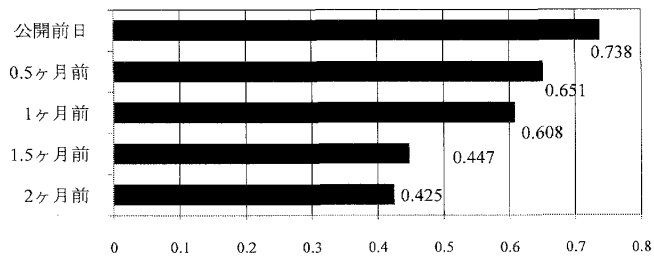

図 7　自由度調整済み決定係数の推移
Fig. 7　Transition of adjusted R^2.

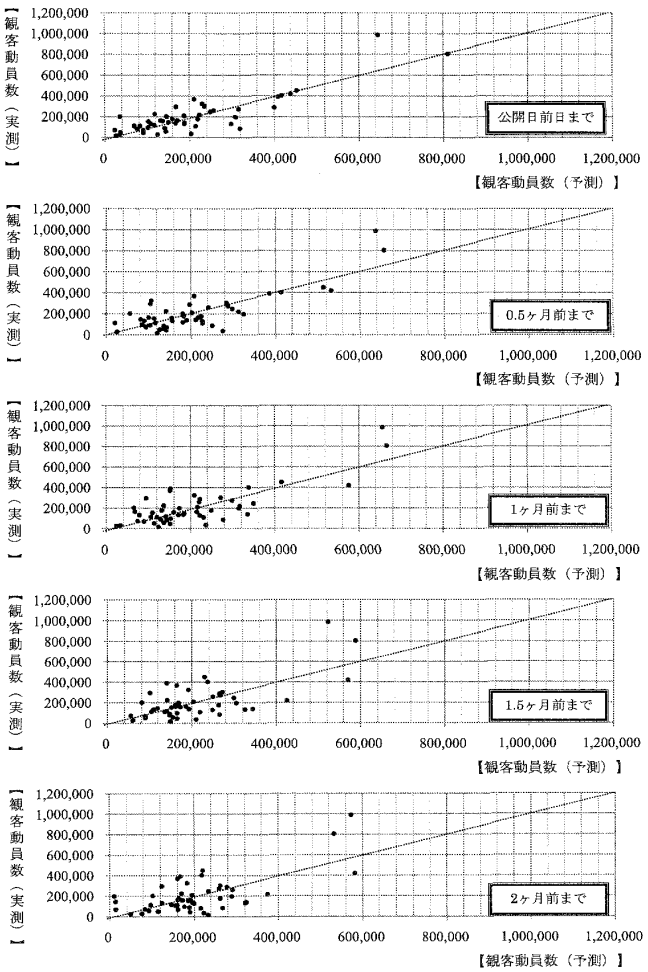

図 8　観客動員数予測の推移
Fig. 8　Transition of movie attendance forecast.

表 7　基本統計量（観客動員数）
Table 7　Descriptive statistics. (attendance)

観客動員数	
平均	203,616.08
標準誤差	25,594.02
中央値	156,934.50
最小	17,480
最大	987,384
標本数	50

B：予測と実測の差異が 2.5 万人以上 5 万人未満，または 25%未満

C：予測と実測の差異が 5 万人以上 7.5 万人未満，または 37.5%未満

D：予測と実測の差異が 7.5 万人以上 10 万人未満，または 50%未満

E：予測と実測の差異が 10 万人以上 12.5 万人未満，または 67.5%未満

F：予測と実測の差異が 12.5 万人以上，または 67.5%以上

この 6 分類における期間ごとの作品数と全体に対する占有率（分類内の作品数/全作品数）が，どのように推移するのかについて示したものが表 8 である．この結果から本研究における提案手法では，映画公開の 1 か月前において観客動員数の予実の差異が 7.5 万人未満または 37.5%未満の範囲（表 8 の A+B+C）で 64%が予測可能であり，実務にも有用な精度をもつことが分かった．

5.2　予測精度に対する TTL 値の影響考察

本節では，**5.1** で示した予測精度に対し，TTL 値がどの程度の影響を与えているかを考察する．

まず，本研究で取り上げた 50 作品の映画公式サイトにおける，TTL 値別の作品数割合を図 9 に示す．TTL 値は，120～259200 秒まで広い範囲で分布している．

図 10 は，観客動員数が同程度で，TTL 値が異なる映画公式サイトにおいて，同一 IP アドレスからの DNS クエリ数を調べた事例である．縦軸は DNS クエリ数であり，横軸は DNS クエリの初観測日から映画公開日前日までのクエリ数を 1 日単位で計測し，DNS クエリ数の多い日順に並べたものである．TTL 値は 900 秒，3600 秒，28800 秒，86400 秒で，観客動員数は，各々，260,115 人，247,042 人，302,068 人，288,050 人である．図 10 に示すように，観客動員数が同程度であっても，TTL 値の違いによらず，クエリ数は多い日で 50 回程度を記録しグラフの傾向も類似している．このことから，TTL 値が短ければクエリ数は多くなるなど，TTL 値の長短とクエリ数の大小には必ずしも明確な規則性があるとはいえないが，本傾向を示す一つの仮説として，同一 IP アドレスから TTL 値間隔以上にクエリ数が発生している一つの要因として，負荷分散装置を設置し，その配下に複数の DNS サーバを配備するような運用を行っていることなどが考えられる．

次に，TTL 値の差異による，各変数（クエリ数，IP アドレス数，国数）と観客動員数との相関係数比較を示したものが図 11 である．TTL 値が 120 秒と 259200 秒については，各々1 作品のため相関係数の算

表 8　作品数と占有率の推移
Table 8　Transition of works number and share.

国数+IP アドレス異なり数の累計	作品数					
	A	B	C	D	E	F
公開前日	20	11	7	2	4	6
0.5 ヶ月前	16	4	14	6	3	7
1 ヶ月前	10	11	11	6	4	8
1.5 ヶ月前	15	6	8	2	6	13
2 ヶ月前	11	9	6	5	1	18

国数+IP アドレス異なり数の累計	占有率		
	A	A+B	A+B+C
公開前日	0.40	0.62	0.76
0.5 ヶ月前	0.32	0.40	0.68
1 ヶ月前	0.20	0.42	0.64
1.5 ヶ月前	0.30	0.42	0.58
2 ヶ月前	0.22	0.40	0.52

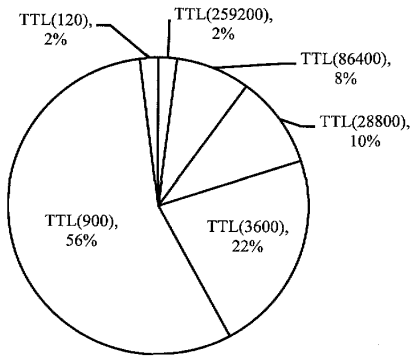

図 9　TTL 値の割合
Fig. 9　TTL ratio.

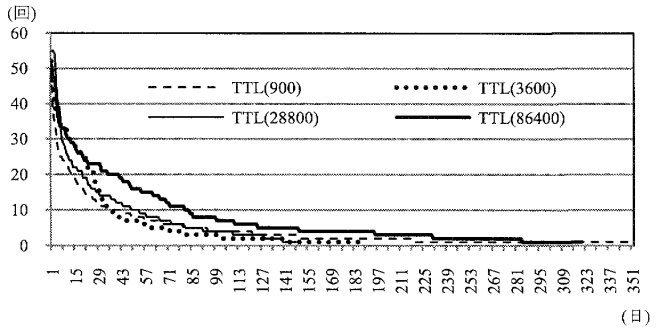

図 10 TTL 値とクエリ数
Fig. 10 TTL and query number.

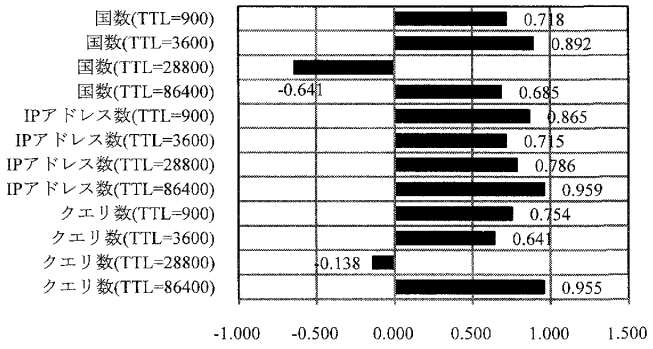

図 11 TTL 値別各変数と観客動員数との相関係数比較
Fig. 11 Correlation coefficient comparison according to TTL.

出は行っていない．図 11 に示すように，TTL 値の長短と観客動員数との相関係数における明確な規則性は見られない．

表 9 は，ある同一の DNS クエリ送出元 IP アドレスにおける TTL 値別，観客動員数別（100,000 人単位）のクエリ数比較である．観客動員数別に示すクエリ数は，DNS クエリ初観測日から映画公開前日までの期間における 1 日当りの最大数である．TTL 値と観客動員数の接点（列と行の接点）内に対象となる映画作品が複数存在する場合は，その中からクエリ数が最大となるものを記載した．例えば TTL 値が 86400 秒で，観客動員数が 100,000 人以下の作品群において，同じ DNS クエリ送出元 IP アドレスから 1 日最大 36 回のクエリが存在した作品があることを示している．表 9 から，本研究で取り上げた 50 作品の観客動員数別 1 日当りのクエリ数と，TTL 値までキャッシュした場合の 1 日当りの想定クエリ数の間には，明確な規則性は見られない．

以上の考察により，次の 3 点が明らかとなった．

① TTL 値の長短とクエリ数の大小には，必ずしも明確な規則性があるとはいえないこと．

② TTL 値の長短と観客動員数との相関係数における明確な規則性は見られないこと．

③ 観客動員数別 1 日当りのクエリ数と，TTL 値までキャッシュした場合の 1 日当りの想定クエリ数の間には，規則性が見られないこと．

これらの結果から，本研究において TTL 値は，観客動員数予測に大きな影響を与えているとはいえないことが分かった．

表 9 TTL 値別，観客動員数別クエリ数比較
Table 9 Query number comparison according to movie attendance and TTL.

TTL 値（秒）		120	900	3600
観客動員数	0-100,000	-	53	74
	100,001-200,000	153	70	52
	200,001-300,000	-	51	52
	300,001-400,000	-	46	-
	400,001-500,000	-	89	-
	800,001-900,000	-	-	88
	900,001-1000,000	-	39	-
TTL 値までキャッシュした場合の 1 日当たりの想定クエリ数		720	96	24
TTL 値（秒）		28800	86400	259200
観客動員数	0-100,000	-	36	-
	100,001-200,000	21	51	15
	200,001-300,000	21	51	-
	300,001-400,000	78	-	-
	400,001-500,000	-	-	-
	800,001-900,000	-	-	-
	900,001-1000,000	-	-	-
TTL 値までキャッシュした場合の 1 日当たりの想定クエリ数		3	1	0.3

6. むすび

本研究では「書き手側」ではなく，意図的な情報操作が入りにくい「読み手側」からの関心度測定手法を提案した．**3.** で，DNS クエリデータから導いた異なり数を用いた「読み手側」からの関心度測定手法の提案を行い，**4.** で，提案手法についての事例検証として日本映画の公開初週の観客動員数予測を行った．提案手法は，従来手法 [15] が映画公開前で相関係数 0.454〜0.542 であったのに対し，0.865 と予測精度において優れている．更に **5.** で，この結果が実務にも適用可能であることを示した．

本論文では DNS という技術が現実社会の分析に利

用できることを示した．本論文では，日本映画を分析対象としたが，その他の事例に対する適用可能性の検討や更なる予測精度の改善（例えば，Web サイトのアクセスログ解析との組合せやブログなど書込み内容を分析する「書き手側」の手法との組合せ，世代や性別など映画作品固有の特性との組合せなど）は，今後の課題である．

謝辞　本研究を進めるにあたり，貴重なアドバイスを頂いたギャガ（株）取締役星野有香氏，及び関係者の方々に感謝致します．

文　献

[1] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins, "The predictive power of online chatter," Proc. WWW, 2007.

[2] 上原　宏，佐藤忠彦，吉田健一，"インターネット・コミュニティ・データを使ったテレビ CM の商品イメージ形成効果測定，" 人工知能誌，vol.23, no.3, pp.205–216, 2008.

[3] P. Kolari, A. Joshi, and T. Finin, "Characterizing the splogosphere," Proc. WWW 2006 3rd Annual Workshop on the Weblogging Ecosystem, Aggregation, Analysis and Dynamics, 2006.

[4] Y.-R. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B.L. Tseng, "Splog detection using self-similarity analysis on blog temporal dynamics," Proc. 3rd International Workshop on Adversarial Information Retrieval on the Web, AIRWeb '07, pp.1–8, 2007.

[5] S.L. Brown, A. Tilton, and D.M. Woodside, "The case for on-line communities," McKinsey Quarterly, vol.1, Retrieved 01, Oct. 2004.

[6] A.M. Muniz, Jr. and T.C. O'guinn, "Brand community," J. Consumer Research, vol.27, pp.412–432, 2001.

[7] R. Tumarkin and R.F. Whitelaw, "News or noise? Internet postings and stock prices," Financial Analysts J., vol.57, no.3, pp.41–51, 2001.

[8] 小川美香子，佐々木裕一，津田博史，吉松徹郎，國領二郎，"黙って読んでいる人達（ROM）の情報伝播行動とその購買への影響，" マーケティングジャーナル，no.88, pp.39–51, 2003.

[9] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," Proc. WWW Conf., pp.491–501, 2004.

[10] 福原知宏，中川裕志，西田豊明，"時系列テキスト集合からの社会的関心の分析，" 第 16 回インテリジェント・システム・シンポジウム，2006.

[11] H. Uehara and K. Yoshida, "Anotating TV drama based on viewer dialogue—Analysis of viewers," Attention Generated on an Internet Bulletin Board-SAINT2005, pp.334–340, 2005.

[12] T. Kurashima, T. Tezuka, and K. Tanaka, "Mining and visualization of visitor experiences from urban blogs," Proc. 17th International Conference on Database and Expert Systems Applications, DEXA, 2006.

[13] C.A.R. Hidalgo, A. Castro, and C. Rodriguez-Sickert, "The effect of social interactions in the primary consumption life cycle of motion pictures," New J. Physics, vol.8, 52, 2006.

[14] L. Yong, "Word of mouth for movies: Its dynamics and impact on box office revenues," J. Marketing, vol.70, pp.74–89, 2006.

[15] G. Mishine and N. Glance, "Predicting movie sales from blogger sentiment," Proc. AAAI Spring Symposium on Computational Approaches to Analysing Weblogs, 2006.

[16] M.S. Sawhney and J. Eliashberg, "A parsimonious model for forecasting gross box-office revenues of motion pictures," Marketing Science, vol.15, no.2, pp.113–131, 1996.

[17] B.R. Litman and H. Ahn, "Predicting financial success of motion pictures," in The Motion Picture Mega-industry, ed. B.R. Litman, Allyn & Bacon Publishing, Boston, MA, 1998.

（平成 22 年 1 月 29 日受付，5 月 10 日再受付）

三田村健史　（正員）

筑波大学ビジネス科学研究科博士後期課程在籍．DNS 及びインターネットドメイン名に関する技術企画，管理運用に従事．(株) 日本レジストリサービス技術研究部長，兼システム運用部長．情報処理学会会員．

吉田　健一　（正員）

1980 東工大・理・情報科学卒，同年日立製作所入社．1992 年 9 月博士（工学，大阪大学）．2002 より筑波大学大学院ビジネス科学研究科教授．インターネット上の各種データを，機械学習の手法を使って解析する研究に従事．情報処理学会，人工知能学会等各会員．