

Doblog の利用に関するアンケート調査からみたユーザ像

Preliminary Experiment on Blog Modernology Project

松村 真宏*¹ 三浦 麻子*²
Naohiro Matsumura Asako Miura

*¹ 大阪大学大学院経済学研究科
Graduate School of Economics, Osaka University

*² 神戸学院大学人文学部
Faculty of Humanities and Sciences, Kobe Gakuin University

In this paper, we introduce our project named 'Blog Modernology' where folkways of people living in the present are to be observed through the blog articles. To conduct this project, web mining techniques are needed for mining blog articles from various point of views, such as gender, living place, events, happiness factor, time of day, day of week, social behavior etc. As a preliminary experiment on Blog Modernology, we here show some simple results obtained from blog articles in Doblog (<http://www.doblog.com/>) in conjunction with questionnaire survey results, i.e., gender-, area-, or age-segmented un/happiness distribution, gender- and area-segmented un/happiness related events, and gender- and age-segmented un/happiness related events.

1. はじめに

ブログサービスは、簡単な手続きで手軽に記事をインターネット上に公開できるサービスであり、昨今急速な勢いで利用者数が増えている。総務省が2006年4月13日に発表した「ブログ及びSNSの登録者数」*¹によると、2006年3月末現在のブログ登録者数は868万人であり、2005年9月末の473万人と比べると半年の間にほぼ倍増している。アクティブユーザ数は割り引いて考える必要があるとしても、ブログが特定の人だけでなく一般の人に定着していることは明らかである。

ブログには個人の日々の出来事が綴られており、ブログを読むと、誰が、どこで、何を、どう感じたのか、といったリアルな生活の断片が伝わってくる。これはいわば膨大な日常生活の記録であり、これまでフィールドワークや文献調査による観察や記録によって行われてきた民俗学や、都市の風俗を観察して研究する考現学[今 87]がブログを通して実現できるようになる。筆者はそのような着想から「ブログ考現学プロジェクト」を立ち上げており、本論文ではその一環として無料のブログレンタルサービスである Doblog*²のブログ記事を「Doblog の利用に関するアンケート」調査*³と併せて分析し、我々の日常に溢れる「幸せ」についてのブログ考現学を試みる。

2. 関連研究

では、ブログ記事からどのようにして著者の(例えば「幸せ」)感情状態を抽出することができるだろうか。ブログサービス LiveJournal*⁴では100種類以上のムードアイコン*⁵の中から自由に1つ選んでブログ記事に付与することができるようになっている。Mishneはこのブログ記事を用いて分類実験を行い、SVMによる分類精度は約50~65%、人手による分類精度は約63%であったと報告している[Mishne 05]。また、MihalceaらもLiveJournalのうちhappyとsadのムードが付与されたブログ記事だけを用いてNaïve Bayes Classifierによるhappy記事とsad記事の分類を行い、約79%の精度

連絡先: 松村真宏, 大阪大学大学院経済学研究科, 〒560-0043

豊中市待兼山町1-7, matumura@econ.osaka-u.ac.jp

*1 http://www.soumu.go.jp/s-news/2006/060413_2.html

*2 <http://www.doblog.com/>

*3 <http://www.team1mile.com/asarin/research/doblog/>

*4 <http://www.livejournal.com/>

*5 <http://www.livejournal.com/moodlist.bml>

で分類できたと報告している[Mihalcea 06]。しかし、happyもしくはsadのムードの付与された限定されたブログ記事だけを分類対象としているため、Mishneによる実験と比較すると問題設定が易しいといえる。1000以上もの感情語に人手でPAD(pleasure, arousal, dominance)のアノテーションが付与されたANEWコーパス[Bradley 99]もあるが、LiveJournalのブログ記事から得られたhappyに関わる語との相関は弱いことが報告されている[Mihalcea 06]。

このように「幸せ」と言ってもその意味は幅広く、その基準も人によってさまざまであるため、ブログの記事内容が幸せに関連する出来事について綴っているかどうかを判別することは簡単ではない。また、LiveJournalには日本語で書かれたブログ記事が少ないため、日本語のムード分類コーパスとして用いるには不十分である。そこで本研究では、幸せな内容の文章を書くときによく用いられる語のリスト、不幸せな内容の文章を書くときによく用いられる語のリストをそれぞれ人手で作成し、そのリストに基づいて幸せ・不幸せの分布および関連イベントの分析を行う。Mihalceaは分類に効いている素性(語)を利用してhappinessの時間分布や曜日分布、happinessと行動の社会性との関係など、happinessの起こる要因に目を向けて分析しており本研究の目的と非常に近いが[Mihalcea 06]、ムードが付与されたコーパスを用いない点、日本語で書かれたブログを対象としている点が異なる。

3. 分析データ

「Doblog の利用に関するアンケート」調査はDoblogユーザを対象に2005年4月22日~5月23日にかけて行われ、758人から回答を得た。アンケートは全部で106問からなり、デモグラフィック属性から利用者意識まで幅広く調査している。その中から本研究では、性別(男性、女性)、年齢(17才以下、18~20才、21~24才、25~29才、30~34才、35~44才、45~54才、55才以上)、住所(北海道・東北、東京都、東京以外の関東、中部・甲信越、近畿、中国・四国、九州・沖縄、海外)の3つの設問結果を利用する。

本研究で対象とするブログ記事は、上記アンケートに答えたDoblogユーザが2003年11月4日から2005年6月27日までに書いたブログ記事約24万件である。ブログ記事の内訳を表1に示す。人数に関しては男性の方が女性より1.8倍ほど多いが、一人あたりの平均記事数および1記事あたりの平均

文字数はほぼ同数であった。

表 1: ブログ記事の内訳

	人数	記事数	平均記事数	平均文字数
男	486	158,535	326	344
女	272	83,829	308	337

4. 幸せ語と不幸せ語の抽出

本研究では、幸せな内容の文章を書くときによく用いられるであろう語を「幸せ語」、不幸せな内容の文章を書くときによく用いられるであろう語を「不幸せ語」とし、それらの語を手がかりとして分析を進める。そこでまず幸せ語と不幸せ語のリストを用意する必要があるが、「positive words co-occur more than expected by chance, so do negative words」の仮説 [Yu 03] が幸せ語と不幸せ語にも当てはまると仮定し、幸せ語のリスト、不幸せ語のリストを手動で作成した。以下に幸せ語のリストの作成手順を示す。不幸せ語のリストも同様の手順で作成した。

Step1) 「嬉しい」「楽しい」などの幸せな内容を表す語を「幸せ語リスト」に入れる。

Step2) Doblog のブログ記事からブログ 1 万記事をランダムに選び、同じ記事中で幸せ語と共起する語（形容詞、サ変名刺、動詞）をカウントし、頻度 10 以上の語を出力。なお、形態素解析には Juman 5.1^{*6} を用い、形態素の表記には「代表表記」を用いて表記の揺れを吸収した。

Step3) 出力された語の中から「幸せ語」に相当するであろう語を手で選択して幸せ語のリストに追加し、Step1) に戻る。追加する語がなくなれば手順終了。

最終的には幸せ語 23 種と不幸せ語 48 種が得られた。

抽出された幸せ語リスト、不幸せ語リストの評価として、(^.^) (^o^)(.^.) などの笑顔の顔文字を含むブログ記事を幸せな記事、(TT) (T_T) (> <) などの泣き顔の顔文字を含むブログ記事を不幸せな記事の正解集合とみなしたときの precision と recall を表 2 に示す。Recall は 4 割弱と低いが、Precision は幸せ語で 5 割弱、不幸せ語で 8 割であり、問題の難しさ [Mishne 05] を考えると許容できる精度であろう。

表 2: 幸せ語・不幸せ語の precision と recall

	幸せ語	不幸せ語
Precision	0.49(35/72)	0.80(39/49)
Recall	0.37(35/94)	0.39(39/101)

ここで抽出した幸せ語、不幸せ語のリストを用いてブログ記事ごとの「幸せ言及度」「不幸せ言及度」を定義する。本研究では、幸せ語の 1 記事中の出現頻度之和をブログ記事の幸せ言及度、不幸せ語の 1 記事中の出現頻度之和をブログ記事の不幸せ言及度とする。幸せ語のリストを L_h 、不幸せ語のリストを L_{uh} 、ブログ記事 x における語 w の頻度を $freq(x, w)$ とすると、ブログ記事 x の幸せ言及度 $h(x)$ 、不幸せ言及度 $uh(x)$ を以下の式で算出する。

$$h(x) = \sum_{w \in L_h} freq(x, w), \quad uh(x) = \sum_{w \in L_{uh}} freq(x, w) \quad (1)$$

*6 <http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html>

5. 幸せの分布

5.1 男女別幸せの分布

ブログ記事ごとに算出した $h(x)$ と $uh(x)$ を男女別に集計した結果を表 3 に示す。女性の方が男性より幸せ言及度・不幸せ言及度の高いブログ記事を書いていることが見て取れる。

表 3: 男女別幸せ言及度・不幸せ言及度

	幸せ言及度の平均	不幸せ言及度の平均
男	0.703	0.439
女	0.916	0.567

5.2 地域別幸せの分布

幸せ言及度の平均・不幸せ言及度の平均を地域別に算出した結果を表 4、日本地図上での分布を図 1 に示す。東京を含む関東、中部・甲信越、中国・四国で幸せ言及度が高く、九州・沖縄の不幸せ言及度が低いことが分かる。なお、男女別にみるとどのエリアでも男性より女性の方が幸せ言及度・不幸せ言及度ともに高かったが、表 4 には男女を合わせた結果のみ示す。

表 4: 地域別幸せ言及度・不幸せ言及度

	人数	幸せ言及度の平均	不幸せ言及度の平均
北海道・東北	71	0.681	0.450
東京都	183	0.841	0.511
東京以外の関東	223	0.804	0.493
中部・甲信越	84	0.815	0.501
近畿	100	0.664	0.466
中国・四国	41	0.823	0.467
九州・沖縄	39	0.650	0.393

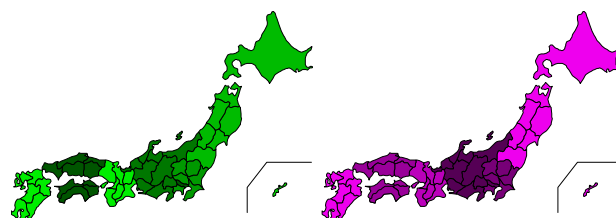


図 1: 左図は幸せ言及度の分布、右図は不幸せ言及度の分布を表す。色の濃いエリアほど幸せ言及度 / 不幸せ言及度が高い。

5.3 年代別幸せの分布

次に、幸せ言及度・不幸せ言及度を年代別に分けて集計した結果を表 5、チャートによる顔グラフによる幸せ言及度・不幸せ言及度の分布を図 2 に示す。これより、基本的に若いときほど幸せ言及度・不幸せ言及度はどちらも高くなる傾向にあるが、18-20 歳の間だけ幸せ言及度・不幸せ言及度が一時的に低くなっていることが分かる。なお、どの年代でも女性の方が男性より幸せ言及度・不幸せ言及度ともに高かった。表 5 には男女を合わせた結果のみ示す。

6. 幸せイベント

6.1 幸せ・不幸せに関連するイベントの抽出

前章での分析によって幸せ言及度・不幸せ言及度の分布が男女別、地域別、年代別のそれぞれにおいて偏りが生じていることが明らかになったが、その要因までは分からなかった。そこで本章では、幸せ・不幸せの要因となるイベントを抽出することによって幸せ・不幸せの要因を探る。

表 5: 年代別幸せ言及度・不幸せ言及度の分布

人	幸せ言及度の平均	不幸せ言及度の平均
17 歳以下	37	0.898
18-20 歳	69	0.797
21-24 歳	132	0.895
25-29 歳	164	0.881
30-34 歳	151	0.798
35-44 歳	147	0.639
45-54 歳	39	0.709



図 2: チャーノフの顔グラフによる年齢別幸せ顔・不幸せ顔の分布。上段が幸せ顔，下段が不幸せ顔である。

ここで，幸せ語・不幸せ語と共に用いられる語が幸せイベント・不幸せイベントを表していると考え，語 w (名詞とサ変名詞) の幸せ言及度 $h'(w)$ ，不幸せ言及度 $uh'(w)$ を以下の式で算出する。

$$h'(w) = \sum_{x \text{ including } w} h(x) \quad (2)$$

$$uh'(w) = \sum_{x \text{ including } w} uh(x) \quad (3)$$

$h(w)$ および $uh(w)$ の高い語がそれぞれ幸せに関連したイベント，不幸せに関連したイベントとなる。

6.2 男女別幸せイベント

5.1 で明らかになった幸せ言及度・不幸せ言及度の男女差の要因を探るために，男女別に集計した幸せイベント，不幸せイベントを表 6 に示す。なお，表 6 は $h'(w)$ の男女比，および $uh'(w)$ の男女比によるランキングの上位 10 語を表示している。これより，女性では「恋愛」が幸せと関係している一方で「結婚」が不幸せと関係していることなどが分かる。

表 6: 男女別幸せイベント・不幸せイベント

順位	男		女	
	幸せイベント	不幸せイベント	幸せイベント	不幸せイベント
1	意味 (1.79)	選択 (1.71)	仕事 (0.79)	仕事 (0.80)
2	作業 (1.76)	記録 (1.70)	心配 (0.72)	結婚 (0.77)
3	追加 (1.76)	連絡 (1.69)	恋愛 (0.72)	生活 (0.77)
4	スタート (1.75)	追加 (1.68)	予定 (0.72)	関係 (0.74)
5	連絡 (1.75)	スタート (1.68)	会話 (0.72)	電話 (0.73)

6.3 地域別幸せイベント

5.2 で示した地域別の幸せ言及度・不幸せ言及度の要因を探るために，地域別に算出した幸せイベント・不幸せイベントを表 7 に示す。しかしながら，今回の分析では地域ごとのイベントの傾向に顕著な違いは見られなかった。

6.4 年代別幸せイベント

次に 5.3 での年代別幸せ言及度・不幸せ言及度の $h'(w)$ ， $uh'(w)$ によるランキングを表 8 に示す。年齢とともに幸せイベント・不幸せイベントが変わっており，20 代前半までは「勉強」や「バイト」といった学校に関連する話題が上位にくるのに対し，20 代中頃以降は「仕事」「結婚」「電話」「生活」といった仕事や家庭に関する話題が上がってくるのが分かる。

7. まとめと今後の課題

本論文では「ブログ考現学」への第一歩として，ブログ記事から幸せ・不幸せの分布や関連イベントの抽出に関する基礎的な分析に取り組んだ。幸せ言及度・不幸せ言及度の計測やイベント抽出については，機械学習による素性抽出 [Mihalcea 06] やモダリティに注目したイベント抽出 [野呂 05] などの従来研究で得られた知見を取り込みながら精度の向上に努めていきたい。また，ブログ考現学は，上記に挙げた課題の他にもイベント発生地域の推定 [安田 06]，イベント発生時刻の推定推定 [野呂 05]，感情推定 [Mishne 05, Mihalcea 06] などさまざまな要素技術の上に成り立っている。これらの関連技術にも取り組みながら，Daily Life Experience [Kahneman 04] をブログから蓄積し，ブログを通して都市風俗を理解していきたい。

謝辞

Doblog の記事データおよび「Doblog の利用に関するアンケート」の調査データは株式会社 NTT データおよび株式会社 ホットリンクより提供を受けました。記して感謝致します。

参考文献

- [Bradley 99] Margaret M. Bradley, Peter J. Lang: Affective norms for English words (ANEW): Instruction manual and affective ratings, Technical Report No. C-1, The Center for Research in Psychophysiology, University of Florida, 1999.
- [Yu 03] Hong Yu, Vasileios Hatzivassiloglou: Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences, Empirical Methods in Natural Language Processing (EMNLP-2003), pp. 129–136, 2003.
- [Kahneman 04] Daniel Kahneman, Alan B. Krueger, David A. Schkade, Norbert Schwarz, Arthur A. Stone: A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method, Science, Vol. 306, no. 5702, pp. 1776–1780, 2004.
- [今 87] 今和次郎: 考現学入門, ちくま文庫, 1987.
- [McCurley 01] McCurley, K. S.: Geospatial mapping and navigation of the web, World Wide Web, pp. 221–229, 2001.
- [Mihalcea 06] Rada Mihalcea, Hugo Liu: A Corpus-based Approach to Finding Happiness, in AAAI Spring Symposium on Computational Approaches to Weblogs, 2006.
- [Mishne 05] Gilad Mishne: Experiments with Mood Classification in Blog Posts, in the 1st Workshop on Stylistic Analysis Of Text For Information Access, 2005.
- [野呂 05] 野呂太一, 乾孝司, 高村大也, 奥村学: イベントの生起時間常判定, 情報処理学会研究報告, 自然言語処理研究会 2005-NL-170, pp. 7–14, 2005.
- [安田 06] 安田宜仁, 平尾努, 鈴木潤, 磯崎秀樹: ブログ作者の居住域の推定言語処理学会第 12 回年次大会論文集 pp.512–515, 2006.

表 7: 地域別の幸イベントと不幸イベント (上位 10 語)

北海道・東北					東京都				
順位	男		女		順位	男		女	
	幸イベント	不幸イベント	幸イベント	不幸イベント		幸イベント	不幸イベント	幸イベント	不幸イベント
1	仕事 (608)	仕事 (608)	仕事 (1531)	仕事 (1531)	1	仕事 (4602)	仕事 (4602)	仕事 (2580)	仕事 (2580)
2	意味 (450)	生活 (414)	結婚 (1299)	結婚 (1299)	2	意味 (3923)	意味 (3923)	意味 (1583)	電話 (977)
3	生活 (414)	意味 (450)	生活 (994)	生活 (994)	3	関係 (2860)	関係 (2860)	ビックリ (1370)	意味 (1583)
4	紹介 (399)	関係 (379)	意味 (866)	意味 (866)	4	期待 (2541)	期待 (2541)	予定 (1066)	生活 (1017)
5	関係 (379)	期待 (263)	関係 (742)	関係 (742)	5	生活 (2221)	電話 (1655)	苦笑 (1037)	ビックリ (1370)

東京以外の関東					中部・甲信越				
順位	男		女		順位	男		女	
	幸イベント	不幸イベント	幸イベント	不幸イベント		幸イベント	不幸イベント	幸イベント	不幸イベント
1	意味 (3067)	仕事 (2954)	仕事 (2782)	仕事 (2782)	1	試合 (1949)	試合 (1949)	仕事 (935)	仕事 (935)
2	仕事 (2954)	意味 (3067)	意味 (1984)	意味 (1984)	2	意味 (1120)	仕事 (1041)	苦笑 (796)	苦笑 (796)
3	期待 (2614)	試合 (2478)	予定 (1632)	電話 (1435)	3	仕事 (1041)	意味 (1120)	結婚 (740)	結婚 (740)
4	試合 (2478)	期待 (2614)	関係 (1511)	予定 (1632)	4	期待 (923)	予定 (885)	意味 (589)	電話 (373)
5	終了 (2093)	終了 (2093)	電話 (1435)	関係 (1511)	5	予定 (885)	練習 (744)	予定 (549)	生活 (464)

近畿					中国・四国				
順位	男		女		順位	男		女	
	幸イベント	不幸イベント	幸イベント	不幸イベント		幸イベント	不幸イベント	幸イベント	不幸イベント
1	意味 (1878)	意味 (1878)	仕事 (984)	仕事 (984)	1	演奏 (601)	勉強 (504)	仕事 (429)	仕事 (429)
2	仕事 (1698)	仕事 (1698)	意味 (934)	意味 (934)	2	勉強 (504)	仕事 (503)	生活 (201)	電話 (191)
3	期待 (1032)	関係 (1002)	生活 (816)	生活 (816)	3	仕事 (503)	意味 (501)	結婚 (198)	生活 (201)
4	関係 (1002)	説明 (858)	関係 (693)	関係 (693)	4	意味 (501)	生活 (427)	ビックリ (196)	意味 (134)
5	火傷 (1001)	火傷 (1001)	勉強 (583)	予定 (535)	5	録音 (495)	存在 (257)	電話 (191)	コメント (152)

九州・沖縄				
順位	男		女	
	幸イベント	不幸イベント	幸イベント	不幸イベント
1	仕事 (838)	仕事 (838)	仕事 (411)	仕事 (411)
2	意味 (654)	意味 (654)	勉強 (275)	勉強 (207)
3	紹介 (440)	関係 (398)	勉強 (207)	意味 (275)
4	関係 (398)	評価 (363)	関係 (172)	関係 (172)
5	評価 (363)	電話 (222)	遭遇 (172)	遭遇 (172)

表 8: 年代別の幸イベントと不幸イベント (上位 10 語)

17 歳以下					18-20 歳				
順位	男		女		順位	男		女	
	幸イベント	不幸イベント	幸イベント	不幸イベント		幸イベント	不幸イベント	幸イベント	不幸イベント
1	入手 (298)	勉強 (325)	意味 (344)	関係 (172)	1	意味 (684)	意味 (553)	意味 (430)	勉強 (308)
2	勉強 (256)	テスト (299)	勉強 (272)	テスト (165)	2	意味 (542)	期待 (450)	勉強 (378)	バイト (249)
3	意味 (254)	授業 (196)	テスト (262)	勉強 (162)	3	期待 (541)	勉強 (416)	授業 (323)	生活 (225)
4	テスト (218)	意味 (191)	関係 (259)	意味 (150)	4	勉強 (531)	バイト (372)	生活 (320)	意味 (220)
5	関係 (188)	入手 (173)	タイプ (256)	授業 (133)	5	バイト (473)	更新 (344)	バイト (298)	予定 (202)

21-24 歳					25-29 歳				
順位	男		女		順位	男		女	
	幸イベント	不幸イベント	幸イベント	不幸イベント		幸イベント	不幸イベント	幸イベント	不幸イベント
1	試合 (2503)	試合 (1385)	バイト (1312)	バイト (849)	1	仕事 (3448)	仕事 (2646)	仕事 (3215)	仕事 (2679)
2	意味 (1886)	意味 (1058)	意味 (1206)	意味 (781)	2	意味 (3245)	意味 (1893)	意味 (1731)	電話 (1398)
3	期待 (1801)	仕事 (902)	仕事 (1135)	仕事 (775)	3	期待 (2580)	期待 (1358)	結婚 (1436)	意味 (1298)
4	仕事 (1289)	期待 (871)	ビックリ (931)	勉強 (659)	4	関係 (2103)	終了 (1183)	電話 (1418)	関係 (1051)
5	予定 (1140)	生活 (733)	勉強 (930)	電話 (647)	5	終了 (2019)	関係 (1168)	関係 (1389)	生活 (983)

30-34 歳					35-44 歳				
順位	男		女		順位	男		女	
	幸イベント	不幸イベント	幸イベント	不幸イベント		幸イベント	不幸イベント	幸イベント	不幸イベント
1	仕事 (3171)	仕事 (2323)	仕事 (3383)	仕事 (2277)	1	仕事 (2691)	仕事 (1889)	仕事 (1733)	仕事 (1233)
2	意味 (2302)	意味 (1465)	意味 (2014)	意味 (1420)	2	意味 (2553)	意味 (1551)	生活 (857)	電話 (519)
3	関連 (1516)	関連 (1048)	結婚 (1880)	苦笑 (1420)	3	関係 (1829)	関係 (1191)	意味 (852)	生活 (495)
4	関係 (1440)	関係 (996)	苦笑 (1762)	結婚 (1292)	4	関連 (1589)	関連 (963)	紹介 (781)	意味 (432)
5	コメント (1244)	電話 (881)	予定 (1586)	生活 (1179)	5	期待 (1405)	電話 (958)	結婚 (692)	予定 (410)

45-54 歳				
順位	男		女	
	幸イベント	不幸イベント	幸イベント	不幸イベント
1	仕事 (1007)	仕事 (587)	仕事 (91)	電話 (114)
2	演奏 (695)	関係 (331)	電話 (64)	仕事 (102)
3	録音 (571)	意味 (297)	関連 (55)	結婚 (49)
4	意味 (542)	生活 (292)	終了 (51)	関係 (46)
5	関係 (485)	原因 (249)	参加 (46)	関連 (46)