視聴滞在時間に基づくWebページ評価手法の検討

Investigation of a Web sites evaluation method based on page-staying time

小池 亜弥[*1] 白井 康之[*1] 小関 悠[*1] 羽野 仁彦[*2] 中川 聖[*2]
Aya Koike Yasuyuki Shirai Yu Koseki Yoshihiko Hano Kiyoshi Nakagawa

[*1](株) 三菱総合研究所 [*2](株) ブログウォッチャー
Mitsubishi Research Institute, Inc. Blogwatcher Inc.,

The "Thatsping" system developed by Blogwatcher Inc., collects visitors search queries and page-staying time from affiliate web sites and web pages. Our objectives are to construct the evaluation rules for web sites according to users' interests which would reflect the page-staying time and to supply the highly evaluated pages according to the queries. One of the analysis methods we use on the "Thatsping" data is a data mining technique called IGVcaep that was constructed through the Inforamtion Grand Voyage Project. The result shows the future possibilities about the dynamic page evaluation based on users' page-staying time.

1. はじめに

近年，インターネットの急速な普及により，膨大な量のWeb閲覧履歴データを入手できるようになってきている．これらのデータはユーザの嗜好や特徴推定のための利用を期待されながらも，実際のビジネスの場においては，確立した分析手法がないため十分に活用されていない．

特に，Webサイトの評価はPageRank[Brin 98]等Webサイトのリンク構造に基づいた手法が提案されているが，それらの手法はユーザの実際の評価や興味を反映していない．

このような背景の中，本論文の目的を次の3点とする．後述のザッピングシステムにより収集したデータ (以下,「ザッピングデータ」) を元にした，(1) ユーザが入力するクエリ間の関連性の発見，(2) ユーザの興味に基づいたWebサイトの評価手法の提案，(3) クエリ毎に評価の高いサイトの発見である．ここでのユーザの興味とはユーザが入力した検索クエリとWebサイトへの視聴滞在時間の組み合わせによって表されるものとする．ザッピングシステムとは (株) ブログウォッチャーが提供するアフィリエートサイトへアクセスした際の検索クエリや視聴滞在時間を収集するシステムである．

ユーザのサイト評価ルール生成には，2008年度経済産業省情報大航海プロジェクト [IGV] にて構築したユーザの特性推定のためのマイニングツールであるIGVcaepを用いた．

また，本論文ではサイト評価に視聴滞在時間を用いることの有効性についても検証する．

2. ザッピングシステム

ザッピングシステム [Thatsping] とは，(株) ブログウォッチャーが提供するJavaScriptコードをブログサイトやウェブサイトに貼付することにより，検索エンジンでの検索キーワードと滞在時間をウェブページ単位で取得したものである．

データに含まれる項目を表1に示す．ここに含まれるレコードIDは，レコード単位で割り振られたシステム上のIDであり，1ユーザの行動を追跡できるものではないため，今回の分析では使用しないこととする．分析対象期間は2008年7月～2008年12月の半年分で，総データ件数は約5，000万件である．

表 1: ザッピングデータ項目

No	項目
1	レコード ID
2	サイト URL
3	サイトタイトル
4	視聴滞在時間（秒）
5	検索クエリ
6	日付
7	時刻

3. 分析方法

3.1 データの絞込み

今回の分析では,「就職活動」関連サイトに焦点を絞って行う．そのため，まず検索クエリに就職活動関連のキーワードを含むもので，データをスクリーニングした．表2はスクリーニングに使用した就職活動関連のキーワードの一例である．

その結果，データは約8万件，ユニークなサイト数は約1万3000件に絞り込まれた．

表 2: スクリーニング用キーワード例

就職活動	就活
適性検査	OB 訪問
面接	新卒採用
筆記試験	会社説明会
SPI	自己分析
志望動機	GAB
エントリーシート	業界研究
内定	自己 PR
インターンシップ	インターン

連絡先: 小池 亜弥
(株) 三菱総合研究所 情報技術研究センター,
東京都千代田区大手町 2-3-6,
a-koike@mri.co.jp

3.2 サイトの特徴付け

全サイトをユニークな存在としてクエリとサイトの二元表を作成した場合，極めて疎なデータになり分析データとして適切ではないため，データの集約化を行う必要がある．そのため，サイト特徴付けルールを作成し，それに基づきサイトの自動分類を行った．表 3 に実際に用いたサイトタグ分類ルールの一例を示す．分析では，サイトの内容に基づき，各サイト特徴付けルールを適用した．サイトによっては複数のサイトタグが付与される場合もある．なお，今回の分析ではサイトタグを 97 種類に分類した．

なお，今回はサイト特徴付けルール作成は人手による作業として実施したが，今後ビジネス化を検討する際には，テキストマイニング技術等を応用することにより自動的な分類を行うことも必要になる．

就職活動関連の 8 万レコードに対して，サイトタグを付与したデータを，サイトタグ付データ D とする．

表 3: サイト特徴付けルール例

タグ	ルール:いずれかを含む
ブログ	blog 投稿 コメント comment trackback トラックバック
掲示板	スレ レス 名無し http://study.milkcafe.net/ http://www.milkcafe.net
ブログまとめサイト	http://bugzero.thatsping.jp/ http://syuukatsu-blog.shooti.jp/
動画サイト	http://img.youtube.com(動画サイト名称)
画像サイト	jpg
面接	面接 質問事項 圧迫面接
エントリーシート	志望動機 履歴書 自己PR フォーマット
自己分析	自己分析
マナー	マナー 礼状例文 スーツ 髪型 服装 封筒の書き方 電話のかけ方
適性検査	適性検査 クレペリン
筆記試験	筆記試験 SPI GAB TOEIC

3.3 関連度分析

本分析では，任意の検索クエリに対して，ユーザの興味が似ている，つまり，関連度の高い検索クエリを見つけることを目的とする．従来のキーワード関連分析に比べ本手法では，ユーザの興味が似通っているという視点からの関連性を見ているため，意味的に近いキーワードのみならず，意味的には必ずしも近くないが興味対象が似通っているキーワードを抽出することができると考えられる．

情報推薦システムの評価方法には精度や再現率等，正確性を示す指標が主に用いられているが，その他に，推薦した情報がユーザに与える有用性や新規性，意外性（セレンディピティ）といった尺度が重要であるとされている [神嶌 07]．一般に，キーワードの共起に基づく方法では、いわゆる “当たり前の結果” しか得られないことが多いが、本稿で対象とするようなユーザの滞在時間に基づく方法では，ユーザが検索したことのないキーワードであってもユーザの興味対象の近いキーワードを抽出できるため，ユーザにとってセレンディピティの高い検索クエリの推薦が期待できる．

データ処理方法は次の通りである．まずサイトタグ付データ D から，検索クエリとサイトタグの平均滞在時間で表される 2 元表データを用意し，これを平均滞在時間行列 A する．クエリ数を m，サイトタグ数を n とすると，平均滞在時間行列 A は $m \times n$ の行列である．A の i 行目をクエリベクトル t_i とする．

また，クエリ i とクエリ j の関連度 $R(i,j)$ は，クエリベクトル t_i，t_j の余弦として以下のように定義される．

$$R(i,j) = \frac{t_i \cdot t_j}{\|t_i\| \cdot \|t_j\|} \tag{1}$$

3.4 サイトの評価ルール生成

3.4.1 データの準備

本分析の目標の一つである，サイトの評価ルール作成を IGVcaep を用いて行う．まずサイトタグ付データ D をサイトタグ毎に分類した．次に滞在時間が 50 秒以上のデータを「クラス H(High)」(評価の高いクラス)，滞在時間が 10 秒以下のデータを「クラス L(Low)」（評価の低いクラス）とした．こうして生成されたデータを IGVcaep を用いて分析した．

3.4.2 IGVcaep によるサイトの評価ルール作成

IGVcaep は，各クラスを特徴付ける顕在パターン（Emerging Patterns）を基にした分類器 CAEP (Classification by Aggregating Emerging Patterns)[Dong 99] に基づき実装されている．CAEP では，特に、アイテム数が多くかつ疎なデータに対して，従来の決定木等に基づく分類器に比べて優れた分類性能が報告されているが，その一方で、各クラスに顕在するパターンを抽出することが計算上のボトルネックとなる．これに対して，IGVcaep では，顕在パターンの抽出に，高速頻出パタンマイニングアルゴリズムとしてよく知られている LCM[Uno 05] の頻出パタン抽出機能を利用している．LCM では，頻出パタン抽出時に各クラスのデータに対して重み付けを行うことにより，各クラスに特徴的なパターンを自動的に抽出することができる．

IGVcaep では 2 クラスで構成されるデータに対して，各クラスにおける頻出パタンの中の顕在パターンを集約することにより分類モデルを構築する．以下に IGVcaep で実装されたアルゴリズムについて説明する．

IGVcaep では 2 クラス (c_i, c_j) で構成される入力データに対し，初めに LCM による高速パタンの列挙を利用し，以下のような条件を満たすクラス c_i の顕在パタンを列挙する．(クラス c_j についても同様である．)

$$GR_{c_i}(p) = \frac{Supp_{c_i}(p)}{Supp_{c_j}(p)} \geq \text{最小増加率}, (i \neq j) \tag{2}$$

ここで，$Supp_C(p)$ はクラス C におけるパターン p のサポートを表す．最小増加率は 1 以上の値をもつパラメータであり，GR_{c_i} はパターン p のクラス c_j に対するクラス c_i の出現頻度オッズである．

続いてクラス毎の集約スコアを求める．与えられたトランザクション s のクラス C に対する集約スコアは以下のように定義される．

$$Score(s,C) = \sum_{e \subseteq s, e \in E(C)} \frac{GR_C(p)}{GR_C(p)+1} \times Supp_C(p) \tag{3}$$

ここで，$E(C)$ はトレーニングデータに含まれる顕在パターンの集合を表す．

クラスの予測においては，各クラスの集約スコアを比較し高い方のクラスを推定する．実際には，各クラスの顕在パターンの数の不均衡のバランスを取るため，集約スコアを以下の方法で正規化した数値を比較する．

$$Norm\_Score(s,C) = \frac{score(s,C)}{base\_score(C)} \tag{4}$$

ここで，$base\_score(C)$ はトレーニングデータのクラス C における集約スコアのメジアンである．

表 4: クエリ関連度算出例

「エントリーシート」に関連度が高いクエリ

順位	クエリ	関連度
1	志望動機	0.71
2	就職	0.64
3	短所	0.64
4	長所	0.64
5	内容	0.61
6	再就職	0.6
7	ゲーム会社	0.6
8	学歴	0.6
9	面接官	0.6
10	男	0.59
11	面接	0.58
12	就活	0.58
13	資格	0.58
14	就職活動	0.57
15	任天堂	0.57
16	就職率	0.57
17	内定	0.57
18	大学	0.57
19	新卒	0.57
20	有利	0.56
21	画像	0.56
22	就職先	0.55
23	女性	0.55
24	就職力	0.55
25	評判	0.55
26	質問	0.55
27	就職氷河期	0.54
28	AO	0.54
29	2次面接	0.54
30	女子	0.53
31	偏差値	0.53
32	ランキング	0.53
33	マスコミ	0.53
34	2CH	0.52
35	商学部	0.52
36	会社	0.52
37	ニート	0.52
38	二次面接	0.52
39	適性検査	0.52
40	対策	0.52

「筆記試験」に関連度が高いクエリ

順位	クエリ	関連度
1	計算問題	0.91
2	職業訓練校	0.89
3	GATB	0.87
4	ロコミ	0.87
5	リクナビ	0.86
6	医療事務	0.86
7	志望動機の書き方	0.85
8	内定取消	0.84
9	埼玉	0.83
10	説明	0.83
11	製造業	0.83
12	シニアの就職	0.82
13	60歳以後、就職、パート、愛知県	0.82
14	一般事務	0.82
15	就職祝金	0.82
16	試験問題	0.81
17	内定取消し	0.79
18	薬剤師	0.78
19	介護福祉士	0.78
20	作文	0.78
21	志望動機書査シート	0.77
22	一般職業適性検査	0.76
23	新宿	0.76
24	過去	0.75
25	添え状	0.75
26	練習	0.75
27	公務員試験	0.75
28	失業保険	0.74
29	職務経歴書	0.74
30	未経験	0.73
31	販売職	0.72
32	ポリテク	0.72
33	アビリティガーデン	0.72
34	非言語	0.71
35	キャリアプラン	0.71
36	介護	0.71
37	書き方	0.7
38	過去問題	0.7
39	種類	0.69
40	リクルート	0.69

4. 実験結果

4.1 関連度分析

表 4 に検索クエリ「エントリーシート」と「筆記試験」に対する関連度が高い検索クエリを表す．表から読み取れるように,「エントリーシート」については，エントリーシートの具体的な記載項目であると想定される「志望動機」や「長所」「短所」といったキーワードと関連度が高くなっている．一方で「筆記試験」に関しては,「計算問題」や「GATB」「非言語」等の筆記試験関連のキーワードの他,「職業訓練校」「シニアの就職」「ポリテク」「アビリティーガーデン」等，再就職や転職関連のキーワードと関連が高くなっている．これは従来の検索クエリの共起分析と異なり，ユーザの評価に基づいた関連度を算出した結果である．

このようにユーザの視聴滞在時間を用いた関連度分析により，任意の検索クエリに対して，ユーザの興味が似通っているキーワードを抽出することができ，サイトの視聴滞在時間向上のための入力支援サービス等への応用が考えられる．

4.2 IGVcaep によるサイトの評価ルール

表 5 に，IGVcaep により生成されたサイトタグ「ブログ面接」における評価ルールの一部を示す．結果の読み方は，例えば,「自己 PR」という検索クエリは評価の高いクラス（クラス H）の中でのサポート（出現確率）は 0.29 であり，評価が低いクラス（クラス L）に比べて評価の高いクラス（クラス H）では 2.15 倍多くなっている.「自己 PR」というパターンが出現した場合に評価の高くなる（クラス H）確率は 0.68 となる．

4.3 クエリの組み合わせによるサイト評価の推定

更に提案手法の応用方法について述べる．提案手法では，クエリの組み合わせによるサイト評価の向上を見ることが可能である．表 6 に IGVcaep により生成した評価ルールを用いたサイトタグ「ブログ面接」における正規化後スコアの算出例を示す．この結果，単体のクエリでは評価が低かったものの，同時に検索することで評価の高くなる組み合わせが発見できた．例えば,「SPI」というクエリは，単体でクラス L のスコアの方が高く，したがって評価が低いと判定されるが,「SPI，サンプル」とクエリを組み合わせることにより，クラス H のスコアが高くなり評価が向上する．このようなサイトの評価を向上させる検索クエリの組み合わせは，検索クエリの入力支援サービスに活用することができる．

このようにユーザの検索クエリの組み合わせに対する非単調なスコアリングを行うことにより，ユーザの指定したキーワードの組み合わせに対する最適なページの評価やこれに基づく検索ランキング表示への応用が可能であると考えられる．

表 5: サイトタグ「ブログ面接」評価ルール例

クラス	サポート	増加率	事後確率	パターン
評価高	0.29	2.15	0.68	自己PR
評価高	0.01	3.28	0.77	自己PR 例 職務経歴書
評価高	0.01	3.00	0.75	自己PR 書き方 職務経歴書
評価高	0.21	3.38	0.77	自己PR 職務経歴書
評価高	0.01	2.18	0.69	面接 質問項目
評価高	0.01	2.31	0.70	例 質問
評価高	0.02	2.14	0.68	質問 採用面接
評価高	0.21	3.34	0.77	職務経歴書
評価低	0.04	4.02	0.80	面接 礼状
評価低	0.02	2.77	0.73	面接 礼状 例文
評価低	0.02	2.52	0.72	面接 例文
評価低	0.02	2.62	0.72	礼状 書き方
評価低	0.02	4.29	0.81	礼状 転職
評価低	0.01	3.15	0.76	例文 転職
評価低	0.02	2.03	0.67	内定 書き方
評価低	0.01	2.32	0.70	内定礼状
評価低	0.01	3.31	0.77	転職 最終面接
評価低	0.04	5.71	0.85	最終面接

表 6: サイトタグ「ブログ面接」正規化後スコア例

クエリ	クラスHスコア	クラスLスコア	推定クラス
SPI	0.0000	0.0005	L
SPI,サンプル	0.0013	0.0005	H
ニート	0.0000	0.0000	推定不可能
ニート,ポリテク	0.0005	0.0000	H
ニート,再就職	0.0013	0.0000	H
一般職業適性検査	0.0004	0.0000	H
一般職業適性検査,ソフト	0.0007	0.0000	H
就職,採用面接	0.0168	0.0284	L
就職,自己PR,採用面接	0.2147	0.0284	L
就職活動	0.0000	0.0125	L
就職活動,採用面接	0.0168	0.0125	H
派遣	0.0000	0.0031	L
派遣,自己PR	0.1986	0.0031	H
派遣,自己PR,例	0.2077	0.0245	H
集団面接	0.0000	0.0032	L
集団面接,採用基準	0.0061	0.0032	H
面接	0.0000	0.1242	L
面接,職務経歴書	0.1654	0.1242	H
面接,自己PR	0.1991	0.1242	H
面接,自己PR,職務経歴書	0.5289	0.1242	H

5. 視聴滞在時間を用いたサイト評価ルールの有効性の検証

本論文では視聴滞在時間のサイト評価ルールへの適用を提案しているが，本章では視聴滞在時間を用いることの有効性を検証する．具体的には，視聴時間を用いて IGVcaep により「評価が高い」あるいは「評価が低い」と判別されたトランザクションのうち実際に当該サイトを見たトランザクションの平均視聴滞在時間と，視聴滞在時間を用いずに IGVcaep により閲覧するサイトの種類（ブログまたは掲示板）を正しく判別されたトランザクションの平均視聴滞在時間を比較する．

5.1 トレーニングデータ・テストデータの準備，評価ルールの生成

視聴滞在時間を含める場合のトレーニングデータとして，ブログサイトと掲示板サイトに分割し，視聴滞在時間を目的変数とし，トレーニングデータ 1 とトレーニングデータ 2 を準備する．トレーニングデータ 1 からブログサイトの評価ルールを，トレーニングデータ 2 からは掲示板データの評価ルールを生成する．なお，トレーニングデータ 1 は 17,528 件，トレー

ニングデータ 2 は 7,602 件のトランザクションを含む．テストデータには 4,744 件のトランザクションを用いる．

続いて視聴滞在時間を含めない場合について，上記のトレーニングデータ 1，2 を統合し，サイトの種類（ブログサイト，掲示板サイト）を目的変数トレーニングデータを作成した．テストデータには視聴滞在時間を含める場合に用いたテストデータと同じデータを用いた．なお，テストデータの平均視聴滞在時間は 22.08 秒である．

5.2 検証結果

表 7 に視聴滞在時間を含めた場合にサイトの種類ごとに高評価と推定されたトランザクションのうち実際に当該サイトを閲覧したトランザクションの視聴滞在時間の平均を，表 8 に視聴滞在時間を含めない場合にサイトの種類ごとに正しく推定されたものの視聴時間の平均を示す．視聴滞在時間を含めた場合と含めない場合の平均視聴滞在時間に，掲示板サイトに関してはほとんど有意な差はみられなかったものの（37.69 秒と 34.37 秒），ブログサイトについては視聴滞在時間を含めない場合の平均視聴滞在時間は 18.19 秒であるのに対し，視聴滞在時間を含めた場合に高評価と推定されたトランザクションの平均視聴滞在時間は 26.03 秒となり大きく差が現れた．また，「どちらとも低評価」と推定されたトランザクションの平均は 15.15 秒とテストデータの平均視聴滞在時間を大きく下回った．これより，ブログサイトについては視聴滞在時間をサイト評価ルールに含めることが視聴滞在時間の予測に有効であることが分かった．

一方，掲示板サイトで大きな差が現れなかった理由を考察する．視聴滞在時間にはサイトの読み込みやスクローリングの時間などのノイズが含まれている．掲示板サイトではブログサイトと比較して，複数人により様々な情報が記述されており 1 ページに含まれる情報も多い．そのため評価時間ではないノイズを多く含み，ブログサイトと比較して視聴滞在時間が長い．実際にテストデータのブログサイトの平均視聴滞在時間は 19.04 秒であるのに対し，掲示板サイトは 33.18 秒と大きく異なっている．このノイズ時間の影響により，掲示板サイトは今回提案した評価ルールが有効でなかったと考えられる．

これらの結果より，視聴滞在時間をサイト評価ルールの作成に含めることにより，新たなデータに対しても視聴滞在時間を有意に推定可能であることが検証され，視聴滞在時間データの潜在的な有効性を確認できたが，掲示板サイトのような複数人により様々な情報が書き込まれているサイトの分析においては、ノイズ時間を考慮する必要があることが分かった．

表 7: 視聴滞在時間を含めた場合

	正推定数	平均視聴滞在時間 (秒)
ブログサイト高評価	754	26.03
掲示板サイト高評価	91	37.69
どちらとも低評価	1,314	15.15
推定不可能	2,282	-

表 8: 視聴滞在時間を含めない場合

	正推定数	平均視聴滞在時間 (秒)
ブログサイト	2,334	18.19
掲示板サイト	744	34.37
推定不可能	857	-

6. まとめ

本論文では，視聴滞在時間を用いたクエリ間の関連度及びサイト評価ルールの提案を行った．本手法は既存の検索クエリの共起分析に基づいたサイト評価ルールと比較して，ユーザの興味を反映したサイト評価手法である．

分析では，ユーザの興味の似通っているキーワードを関連度分析により算出した．これは，サイトの視聴滞在時間を向上させるための検索クエリ入力支援サービスに応用できると考えられる．既存の入力支援サービスと比較した新規性は，ユーザの評価値に基いて関連度を算出している点である．本分析で得られる結果は，ユーザが検索したことの無いキーワードであっても興味対象範囲の近いキーワードを抽出できるため，ユーザにとってのセレンディピティの向上が期待できる．

また，IGVcaep によるサイトタグ毎のユーザの評価ルール生成も行った．これは，検索クエリ入力支援のみならず，サイトのリコメンド，検索結果の表示ランキングの入れ替え等のサービスにも応用可能であると考えられる．検証結果により視聴滞在時間データをサイト評価ルールに用いることの有効性が示され，将来的には，検索キーワードに対応したダイナミックなウェブページ評価に基づいた広告ビジネスへの応用も考えられる．

一方で，視聴滞在時間に含まれるノイズの扱い方は今後の課題である．ノイズの扱い方は，例えばノイズ時間の分布を仮定し視聴滞在時間から差し引くことが考えられる．また，今回はサイトへのタグ付けを行う際に検索クエリの表記ゆれについては対処していない．今後表記ゆれを統一化し，シソーラスの概念を取り入れることにより，さらに判別力の高いサイト評価ルールを作成していきたい．

謝辞

本研究は平成 20 年度経済産業省情報大航海プロジェクト [IGV] の一環として実施したものである．

参考文献

[Brin 98] S. Brin, *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, Computer Networks and ISDN Systems, pp.107–117, 1998.

[Thatsping] Thatsping system : `http://thatsping.jp`

[IGV] 情報大航海プロジェクト：`http://www.igvpj.jp/`

[Uno 05] T. Uno et al, *LCM ver.3: collaboration of array, bitmap and prefix tree for frequent itemset mining*, Int'l Conference on Knowledge Discovery and Data Mining, pp.77–86, 2005

[Dong 99] G. Dong, X.Zhang, L.Wong, and J.Li,*CAEP: Classification by Aggregating Emerging Patterns*, Proceedings of the 2nd International Conference on Discovery Science,Tokyo, Japan (pp. 30-42). Springer-Verlag, 1999

[神嶌 07] 神嶌 敏弘, 推薦システムのアルゴリズム (1), 人工知能学会誌 22 巻 6 号, pp.826–837, 2007