1B2-4

陸上競技ブログからの活動記録抽出
Extraction of athlete's activities from blog articles

佐野 正和[*1] 福原 知宏[*2] 増田 英孝[*1] 山田剛一[*1]
Masakazu Sano Tomohiro Fukuhara Hidetaka Masuda Koichi Yamada

東京電機大学[*1]
Tokyo Denki University

独立行政法人産業技術総合研究所[*2]
National Institute of Advanced Industrial Science and Technology

The aim of this study is to encourage people to maintain motivation. In this study, we propose a motivation support system for athletes based on information extraction from blog articles of athletes. Because blog articles written by athletes often contain description about practice menus and records in competitions, we consider that such information is valuable and encourage other athletes. We describe how we extract activities from blog articles and show results of extraction experiment by using many blog articles.

1. はじめに

ある目標を達成するための努力を継続するにはモチベーションが高いことが重要である. モチベーションには, 内発的動機づけと外発的動機づけの2種類が存在する. 内発的動機づけは, 好奇心や関心によって動機づけられることであり, 一方, 外発的動機づけは報酬や何らかの目的を達成するために動機づけられることである. ある学校で, 学習面におけるライバルをもつ生徒の実態調査の研究が実施された[1]. 調査によると, 学習面でのライバルを持つ生徒の成績は, ライバルを持たない生徒よりも高いことが報告されている. しかし, 必ずしも闘える, もしくはお互いに意識を高め合うライバルが身近に存在するとは限らない. そこで, 本研究では陸上競技を取り上げ、競技に取り組む選手に対し、陸上競技をテーマとするブログ記事からその練習内容や大会記録を抽出し、選手に提示することで, 選手の外発的動機づけを高めるシステムを開発する.

本研究の目的は, スポーツや勉強といった分野で個人のモチベーション維持を支援するシステムを構築することである. システムの利用者が取り組む分野において他の人々の活動とその結果に関する情報を収集し, 利用者に提示することで, 利用者のモチベーション維持を支援する. 陸上競技選手のブログは、日々の活動を表現する活動内容と実力を表す大会記録が掲載されており, ライバルとして扱うために必要な情報が得られる.

2. システムの概要

本節では, モチベーション維持を支援するシステムの概要および機能について述べる. 本研究で開発するシステムには, 陸上競技短距離分野を専門とするブロガの活動記録を自動で登録する. 登録されているブロガの活動記録を用いて、モチベーションを維持したい利用者に対し, ブロガの実力が自身と対等であればライバルとして意識させ, また, 遥かに実力が高いブロガであれば活動の内容量を明確に提示し, 日々の活動に反映させることで, モチベーションを維持させる. 図1にシステムの提示機能の流れを示す. 図1に示した提示機能の流れにおいて, 矢印に順序を示す番号と役割が付与されている.

連絡先:佐野　正和, 東京電機大学大学院未来科学研究科
東京都足立区千住旭町 5 番, 03-5284-5333
sano@cdl.im.dendai.ac.jp

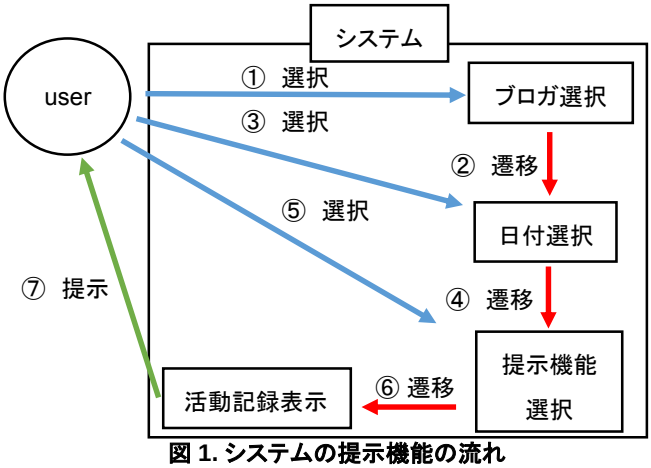

図 1. システムの提示機能の流れ

2.1 データ収集

本システムとは別の抽出システムによって得られたブロガの活動内容と大会記録をシステムに登録する.

2.2 データ入力

システムの利用者は, 他の利用者に提示するため, 自身の行ってきた最近の活動内容と大会記録を入力する.

2.3 ライバルの推薦

大会記録を入力後, 利用者に対して, システムに登録された人物の記録が自身と同等である場合, その人物をライバル候補として推薦する.

2.4 活動記録の提示

特定のブロガおよび日付を選択すると, その日のブロガの活動内容を箇条書きに提示する. 更に,「時系列的表示」および「累計表示」といった提示形式に変更することができる

2.5 活動内容量の時系列的表示

大会前の特定の期間内に, ブロガの行ってきたメニューの内, 陸上競技の短距離分野において重要なトレーニングの内容量を時系列にグラフで可視化する. 図2に時系列に内容量を可視化したグラフの例を示す. AとBは実際のブロガであり, Y軸は1日の走行距離を表している.

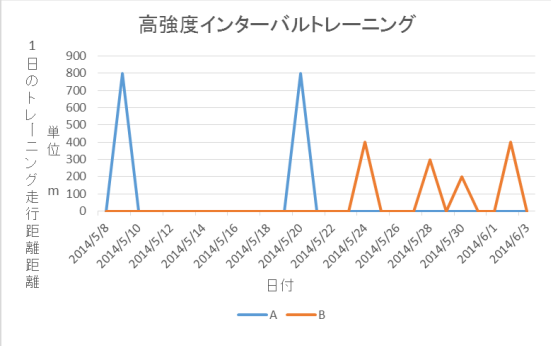

図2. ブロガ2人の活動内容の量を可視化したグラフ

2.6 活動内容量の累計表示

重要なトレーニングの内容量を特定期間内において累計したグラフの可視化を行う. 図3にブロガ2人の活動内容の量を累計したグラフを示す.

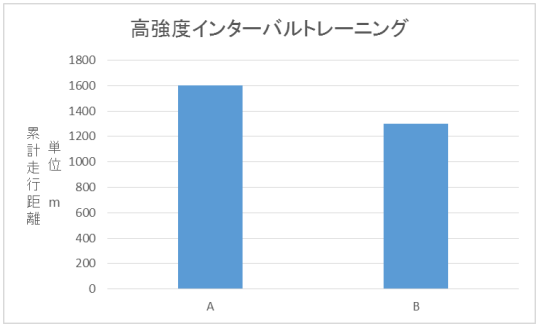

図3. ブロガ2人の活動内容の量の累計を可視化したグラフ

3. 陸上競技ブログ

本節では, システムで提示されるブロガの活動記録の情報源である陸上競技ブログについて説明する. 陸上競技ブログは, 独自に書いた日記のような個人ブログや組織である団体のブログの2種類が存在する. 団体ブログサイトは, 多人数の選手の活動と記録を記載しており, 中には有力選手の情報も含まれている. 2節に分けて, 収集対象である陸上競技選手の活動内容と大会記録を実際の陸上競技ブログの記事を用いて説明する.

3.1 活動内容

陸上競技の活動内容は, 特定の表記であるメニュー名が存在する. 図4に活動内容が掲載されたブログの記事例を示す. ブログの記事中において活動内容の表記は, メニュー名と内容とその内容の量で構成されている.

赤: メニュー名　　**青**: 内容　　**緑**: 量

近所の公園
13:20～15:00
・Jog
・流し100m×10
・坂ダッシュ60m×10
・テンポ走250m×1
・チューブトレーニング
・ランジ20歩×2
・フライングスプリット30×2
・ボックスジャンプ
・down

図4. 活動内容が掲載されたブログ記事

3.2 大会記録

陸上競技の大会記録は, 時間で表記され, また短距離分野においては, 大会記録と同時に風速が計測される. 図5に大会記録が掲載されたブログの記事例を示す. この図で示した赤の下線が引かれた表記が風速である.

60m 7.23(+0.2) M40 1位 ※大会新
100m 11.53(+0.8) M40 1位 ※大会新
200m 24.33(+1.5) M40 1位

図5. 大会記録が掲載されたブログ記事

4. 活動記録抽出手法

3節で説明した陸上競技ブログから活動内容と大会記録の抽出手法を2節に分けて述べる.

4.1 活動内容の抽出手法

ブログ記事から活動内容を4段階に分けて抽出を行う.

(1) 陸上競技練習用語辞書の作成

記事の本文から活動内容を抽出するため, 形態素解析用ツールMeCabの辞書に記事中に頻出するメニュー名を100件追加した. 3.1で示したブログ記事のすべてのメニュー名は, 他のブロガの記事中にも頻出するため, 追加の対象となる.

(2) 正規表現による抽出

3.1節において、ブログに掲載されている活動内容は, メニュー名と内容とその内容の量で構成されていると説明した. 量であるセット数の表記は, 活動内容の手がかりとなるため, 正規表現による抽出を行う. 表1に実際の活動内容を用いて、セット数の表記を手がかりとした正規表現の例を示す.

表1. 正規表現による抽出

演算子	メニュー名
(|.+?)(×\\d)(|.+?)	テンポ走150m×5

(3) 形態素解析による抽出

形態素解析用ツールMeCabを用いてブログ記事を読み込むと, 陸上競技用語辞書を追加したため, 活動名が表記された文を判別する. しかし, 活動内容ではない活動名を含んだ通常の文も抽出してしまう可能性があるため, 活動名の前後の形態素にも着目し, その形態素が文を構成させるための品詞である助詞や動詞であれば, 単なる文章である可能性が高いため抽出の対象外とする.

(4) 活動内容の分解

抽出した活動内容を「メニュー名, 内容, 量」に分解する. 活動内容の量であるセット数の表記は「×」と数字で構成されているのがほとんどなため. 活動内容から量を分離する際, 「×」の表記を用いた正規表現による分解を行う. 表2に分解の例を示す.

表2. 量を抽出する正規表現

リテラル演算子	活動内容
(|.+?)(×\\d)(|.+?)	テンポ走120m×5

↓ **量の抽出**

演算子	内容	量の演算子	量
(|.+?)	テンポ走120m	(×\\d)(|.+?)	×5

量の表記が分離された特定の表記に対してメニュー名とその内容の分解を行う．メニューの内容は，例外はあるが，ほとんどがそのメニューの回数や距離を表現しているため，表記が数字である．従って分離をする際に，通常の文字と数字の境目を利用した正規表現による分解を行う．表3にメニュー名と内容の分解例を示す．

表3.内容を抽出する正規表現

演算子	メニュー
(\\D{2,})(|約|)(\\d{1,})(|.+?)	テンポ走120m

↓ メニュー名と内容の分離

演算子	(\\D{2,})(|約|)	(\\d{1,})(|.+?)
項目	メニュー名	内容
活動内容	テンポ走	120m

4.2 大会記録の抽出手法

本節では，記事本文から大会記録を抽出する手法について述べる．抽出の対象となる記録は，ブログ記事と同じ日付あるいは1日前に出場した大会の記録である．大会記録を抽出する際に，個人ブログか団体ブログかにより抽出手法が異なる．本研究において,陸上競技ブログを25件収集し，分類を行った結果，個人ブログが20件，一方団体ブログは5件であった．しかし団体ブログの件数が少なく，ブログの記事の掲載形式も偏っているため，5件のブログの掲載形式に適応できる手法を提案することができない．よって本研究では、個人ブログのみを対象とする．

(1) 抽出の対象とするブログ記事の限定

大会記録を抽出する際に，練習で計測した記録や過去の記録の抽出を避けるため，対象とする記事を限定する．

(1-1) 活動内容が掲載された記事の除外

練習で計った記録の抽出を避けるため，活動内容が掲載された記事を抽出の対象外とする．収集したブログの執筆者のほとんどが自分の大会について掲載している記事に練習に関する報告をしていない．つまり活動内容が掲載された記事を抽出の対象外としたとき，大会記録が掲載された記事が含まれることはない．

(1-2) 過去の記録がまとめられた記事の除外

収集したブログの執筆者の中には，1年のまとめとして年末やシーズンオフの始期の記事にその年に出した大多数の記録を列挙している．図6に，大多数の記録が列挙されたブログ記事の例を示す．

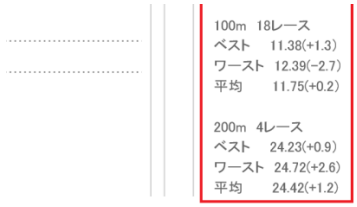

図6. 大多数の記録が掲載されたブログ記事

本研究の抽出対象は，ブロガが出場した大会の日付に掲載された記事の大会記録であるため図6のような記事を抽出の対象外とする．

(2) 他者の記録の排除

個人ブログは過去の記録や他人の記録を掲載することもあり，そのような記録であることを指し示す手がかり表現が，大会記録の掲載された文を含めて前の3行以内にある場合が多く見られた．手がかり表現としては，「彼」「彼女」「男の子」「女の子」「保持」および人名となるものを含めた6種類を使用する．図7に個人ブログの記事において，他者の記録が掲載された例を示す．

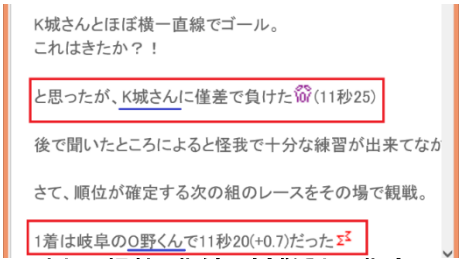

図7. 他者の記録が掲載された記事

(3) 風速を手がかりとした抽出

短距離を専門とするブロガは大会記録と同時に風速も掲載しているので，風速が付与された記録のパターンを表現する正規表現を用いて抽出を行う．表4に風速を手がかりとした正規表現による抽出の例を示す．

表4. 風速を手がかりとした正規表現による抽出

大会記録	演算子
11秒48 (-0.6)	(|.+?)\\d+?秒\\d+?(|.+?)(±0|0|\\+|\\-)\\d\\.\\d(|.+?)

(4) 風速が付与されていない記録の抽出

風速が掲載されていない場合，記録が掲載された文章の近くにブログの記録であることを示唆する手がかり表現である「400」と「結果」を用いて抽出を行う．風速が付与されていない大会記録は種目が400mでありトラックを一周するため，正確な風速を計測することができない．400mの記録を掲載しているブログの執筆者は，大会記録の直前に種目名である「400」を記載している．更に，記事に大会記録を掲載する際に「今日の結果」というように，「結果」という表現も記載しているため，「結果」も抽出の手がかりとなる表現とする．図8に風速が付与されていないブログの記事を示す．

図8. 風速が付与されていない記事

5. 抽出結果の評価

ブログの記事から4節で提案した手法を用いて，適切に活動内容と大会記録を抽出できたかどうかを確かめるため，正解データを用いて，精度と再現率を評価した．収集したブログから無

作為に記事を200件以上抽出し，記事集合に存在する活動内容と大会記録を正解データとする．精度と再現率は式(i)と式(ii)を用いて求める．

$$精度 = \frac{記事中のデータの内正しく抽出した数}{システムが抽出したデータの数} \ldots (i)$$

$$再現率 = \frac{システムが正しく抽出した数}{記事中の正解データの数} \ldots (ii)$$

5.1 活動内容の抽出評価

本節では，活動内容の抽出実験の評価について述べる．人手で用意した正解データを3つの項目に分解し，抽出した活動内容と正解の活動内容を照らし合わせた．評価を行う際には，活動名と内容と量のすべてが正解データと一致するときに正解とする．

ブログ検索エンジンにおいて，20名分の個人ブログを特定し，その内7,029件の記事を取得し，日付が2014年11月10日から同年12月10日までの記事292件から無作為に200件を抽出し，テストデータとして使用した．表5に実験の評価結果を示す．精度は67. 3%，再現率は57. 3%であり，記事200件の中で，見つかった活動内容の総数は1,166件であり，システムが抽出した1,009件の内680件の抽出に成功した．

表5. 活動内容抽出の評価結果

活動内容	不要な抽出データ	合計	精度
680件	329件	1009件	67.3%

抽出	未抽出	合計	再現率
680件	486件	1166件	58.3%

5.2 記録の抽出の評価

5.1と同様に7,029件の記事から，2013年10月1日から同年10月31日までの記事292件をテストデータとして抽出した．抽出した大会記録は，記事中の大会記録が列挙された文であり，評価を行う際には，大会記録が列挙された文を正解データとする．抽出した文と正解の文が一致したとき，正解とする．表6に実験の評価結果を示す．精度は80. 9%，再現率は73. 9%であり，292件の記事中に存在する大会記録は23件であり，システムが抽出した21件の内17件が一致した．

表6. 大会記録抽出の評価結果

大会記録	不要な抽出データ	合計	精度
17件	4件	21件	80.9%

抽出	未抽出	合計	再現率
17件	6件	23件	73.9%

6. 考察

本節では，5節で述べた実験結果の考察について述べる．表7に活動内容の抽出実験で得られた未抽出データの例を挙げ，表8に分解したときの結果を示す．これらの特徴は，記事中に複数の活動内容が1項目にまとまっているものが多く，3つの項目に正確な分解を行うことができなかった．メニュー名と内容に文字と数字の境目を基点にした分離を行った結果，内容はメニュー名が付与された表記で，メニュー名は他の活動内容が付与された表記となった．

表7. 活動内容の不正解データの例

正確に分解できなかった活動内容
・8歩バウンディング→20歩バウンディング ・スキップからの70m流し ・10mスキップからの60m流し×3種

表8. 分解結果

メニュー名	内容	量
バウンディング→20歩バウンディング	8歩	1
スキップからの	70m流し	1
スキップからの60m流し	10m	3

表9に大会記録の抽出実験で得られた不要な抽出データと未抽出データの特徴の例を挙げる．不要な抽出データは，「ベスト」という表記が含まれたブロガの過去の最高記録と「4継」という団体競技の記録が見られた．これらのデータは，ブロガ記事と同じ日付あるいは1日前に出場した大会の記録ではないため，抽出対象ではない．一方，未抽出データの特徴は，風速の手がかり表記である「＋」の記号が付与されていないものや風速が記事中の別の行に付与された記録が見られた。

表9. 大会記録の不正解データの例

不要な抽出データ	未抽出データ
・4継 50秒79 5位 ・ベスト22秒64(+1.8)	・11秒33 (0.1) ・11秒86

7. おわりに

本研究では，モチベーション維持支援システムで扱うための仮想的なライバルの情報を取得するために，陸上競技選手のブログの記事を対象に，活動内容と大会記録の抽出手法を提案し，実際のブログ記事であるテストデータを用いて実験を行った．その結果，活動内容の場合は，精度が67.3%，再現率が57.3%であり，一方，大会記録の場合は，精度が80.9%，再現率が73.9%という結果が得られた．本研究を通して，モチベーションを維持するシステムに必要な材料となる活動記録を実際のブログから抽出した．今後は，システムの提示画面において出力する活動内容の量の可視化を行い，システムの構築に向けてモチベーションが維持できる機能を提案する．

参考文献

[1] 太田信幸，“学習活動におけるライバルの意味−ライバル認知の分析を通して”，名古屋大学教育学部紀要，心理学，46巻，pp.309−310，1999.