SOM を用いたインターネット利用者の性別・年代の推定に関する研究

新名　玄[†1]　櫻田ユカリ[†2]　堂園　浩[†3]　中國　真教[†2]

概要：インターネットの普及とともに，それを活用した広告は市場にとってますます重要なものとなった．しかし，興味・関心を持たない人がその広告をみても意味がない．そのため近年，性別・年代ごとにターゲットを絞ったターゲティング広告が重要視されるようになった．しかし，利用者のログイン中にしかその情報を取得することができないなど，その利用は限られていた．そのため,ログアウト中でも性別・年代を推定する何らかの手法が必要である．そこで，本研究では，ニューラルネットワークのひとつである自己組織化マップ(SOM)を用いて，インターネット利用者の性別・年代を推定する手法について検討を行った．

キーワード：自己組織化マップ，テキストマイニング，ニューラルネットワーク

Self-Organizing Map-based prediction of Internet users age and gender

GEN NIINA[†1]　YUKARI SAKURADA[†2]
HIROSHI DOZONO[†3]　MASANORI NAKAKUNI[†2]

Abstract: With the growth of the Internet, it becomes more and more important for a marketing company to make use of the internet for their advertising. However, to place an advertisement on online to a person who are not interested in the one is not make any sense. Therefore, the targeted advertising focused on the online users age and gender is attracted attention, recently. However, the information can be acquired from a site profile only while the user logs in. Some method for estimating the age and gender also while the user logs out is necessary. Here we propose Self-Organizing Map-based method for predicting Internet users age and gender using the html file on a website.

Keywords: Self-Organizing Map, Text mining, unsupervised learning, Neural Network

1. はじめに

インターネットの普及によって，オンライン上での広告の市場はますます拡大しつつある．しかし，インターネットを利用する人は様々で，それぞれ興味や価値観，趣味嗜好などが異なるため，ただ闇雲に広告をばら撒いているだけでは，その広告に興味関心を持たないユーザにも届くことになり，大変非効率である．それよりも性別・年齢ごとにターゲットを絞り広告を配信する方が，各性別・年代ごとの趣味嗜好の傾向を配慮することができるため，より効率的である．実際に性別・年代を指定してターゲットを絞った広告は効果的で需要も高く，その単価も高い．

しかし，ユーザの性別・年代の情報は，サイトのログイン中におけるユーザ情報から得るなど，ごく限られた場合でしか入手できないため, ログアウト時でもユーザの性別・年代を推定する手法について研究開発することは非常に重要である．しかし，日本でこのような研究開発はまだ進められていないに等しい．日本での性別の推定に関する研究として，クエリログを用いたものがある[1]．この研究ではクエリを形態要素解析にかけるのではなく，性別ごとに作成した文字 *n*-gram 言語モデルを用いた手法を提案している．

我々は，検索クエリではなく，ウェブページにおける html 情報を用いた，ユーザの性別・年代の推定手法について提案する．

2. SOM を用いたインターネット利用者の年代と性別の推定

本研究では，インターネット利用者の性別と年代の推定に利用するデータとして，アクセス先のウェブページに表示されるテキストの頻出ワードを用いた．この頻出ワードは，ユーザが興味関心をもったウェブページの特徴として考えることができ，頻出ワードを用いることで，あるユーザのアクセスするウェブページが，どのような年代や性別から興味関心を持たれやすい特徴を含んでいるかという評

†1 (有)SOM ジャパン
SOM HAPAN Inc.
†2 福岡大学
Fukuoka University
†3 佐賀大学
Saga University

ⓒ2017 Information Processing Society of Japan

価をすることで，そのユーザの性別と年代の推定を行う．具体的には，性別・年代ごとに興味関心のあるウェブページの特徴をあらかじめ，ベクトル化したモデルとして分類し，その基準となるモデルとユーザが実際にアクセスするウェブページの特徴を表したモデルとの比較を行い，どのモデルに最も類似しているかを評価することで，ユーザの年代性別を推定する(図 1).

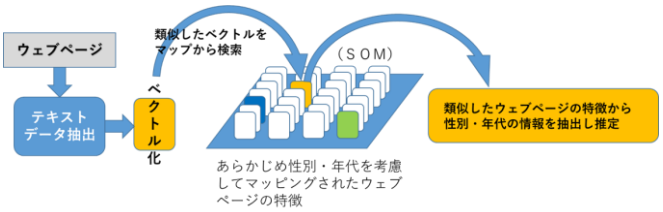

図 1：性別・年代の推定手法に関する概略図

ウェブページの特徴をベクトル化する手法には，頻出ワードを要素としたベクトルを用いており，ウェブページのテキストからこれらの頻出ワードを切り出す作業が必要である．テキストからワードを切り出す方法については，形態要素解析のフリーソフト“chasen”や，“MeCab”などが有名であるが，本研究では独自にワードの頻度を抽出するアルゴリズムの開発をおこなった．
性別・年代ごとのウェブページの特徴を表すベクトル化されたモデルの分類には，高次元ベクトルの分類に有用な自己組織化マップ(Self-Organizing Maps：SOM)[2][3]を用いた．自己組織化マップは，学習で与えられたモデルが分類されるだけでなく，データの数に対してマップのサイズが大きい場合，あるモデルとあるモデルの両方に類似したような新しいモデルを作ることができ，推定の際に，もとのモデルの数以上の様々なモデルパターンと照らし合わせることができるため，より多くのユーザパターンを推定できることが期待される．

以下, 頻出ワードの抽出手法, SOM の原理について述べる．

2.1 ウェブページにおける頻出ワードの抽出手法

本研究では，ウェブページのテキスト情報として，html ファイルを用いる．html ファイルには，ウェブページとして実際に表示される文字とは別に“属性”などの情報が含まれており，そのまま頻出ワードを検索しようとするとテキスト以外の文字列についてもワードの出現頻度をカウントしてしまうため, 意味のない文字列が抽出されてしまう．

したがって，適切なワードを抽出するためには，はじめに html の情報を必要のない文字を削除した意味のある文字のならびに変換しておく必要がある．
つまり, 流れとしてはまず html の情報を意味のあるテキスト情報に変換し，その後，頻出ワードの抽出処理を行うことになる．以下，それぞれの手法について述べる．

2.1.1 html 情報をテキスト情報に変換する手法

実際にウェブページで表示されるテキストは，html ファイル上では，~や ~ ， ~などのタグとよばれる文字列の並びで挟まれており，これらの間に挟まれた情報を抜き出すことでテキスト情報を得ることができるようにみえる．しかしタグは入れ子にして ~~~~~~ という形で記述できることや，その中に直接テキストとは関係のない“属性”と呼ばれるコードが含まれるなど，単に挟まれた部分を切り出すだけでは，意味のない文字列も切り出してしまうことになりかねない．しかし基本的に，切り出しに必要な意味のあるテキスト情報は ”> ～ ＜ や ～<に挟まれていることに注目すれば，この２種類の範囲で挟まれた部分を抜き出すだけでよいため，すべての必要なタグをすべて検索し，入れ子の仕方がどのようになっているのかや，属性はどの部分に含まれているかを解析して抽出するよりも早く抽出が行える．しかし，それでもこのルールには例外があることも事実で，例外を解決する方法として，”> ～ ＜ や ～ ＜ で挟まれている部分を切り出す前に, 次の表 1 に示すような例外的な文字列を空白に変換しておくことで，このルールが適応できる．このような手法を用いて，本研究ではウェブページから必要な情報を切り出している．

表 1：空白に変換すべき文字列の例

		<code>	JavaScript"	javascript"
		</code>	ostyle"	json"
</i>		 	gstyle"	css"
">'	">"	="">	polyfill">	common">

2.1.2 テキスト情報から頻出ワードを抽出する手法

テキストデータから頻出ワードを切り出す手法について，その手順を以下に示す．

STEP 1：

2.1.1 の手法によってウェブページから切り出されたテキストデータの初めからi 番目の文字を $s[i]$とする．また，h (初期値は 1)番目からi 番目までの文字列を取り出し，これを先頭から

$$w[1] = s[h], \quad w[2] = s[h+1], \quad \dots, \quad w[i-h+1] = s[i]$$

となるように，文字列 W を定義する．w[1]とテキストデータにある文字を先頭から比べ，
w[1] と $s[j]$ が 同じ文字になる位置をすべて探す. ただし，抜き出した場所は検索しない．

ⓒ2017 Information Processing Society of Japan

このとき，gの初期値は 1 として，$w[g]$~$w[t]$ (ただし：$1 \leq g < t \leq i-h+1$) と $s[j]$~$s[j+t-g]$がこの順番で連続して同じになるようなすべての t についての文字列$w[1],w[2],..,w[t]$を考える．このような t が存在しない場合は，g の値を 1 つ増やして同じ操作を t が存在するまで繰り返す．tが存在する場合は STEP 2 に移る．

STEP 2：

STEP 1 で得られたすべての t に関する文字列$w[1],w[2],..,w[t]$がテキストデータ内にいくつ存在するかをカウントし，このカウント数をC_t とする．ただし，ただし，$s[h-1]$，$s[h]$ がともにカタカナである場合は，$s[h-1]$のカタカナの種類が 2 つ同じ場合は，カウント$C_t = 0$とする．このとき，$\max_t C_t \times t \times t$ を満たす t で文字列$w[1],w[2],..,w[t]$をテキストから抽出する．ここで，抽出された文字列の出現頻度はC_tである．
次にh を$g+t$ 増やして，STEP 1 にもどる．

このようにして，STEP 1～STEP 2 を繰り返しながら，文字列を抽出していく．
このとき，同時に文字列の出現頻度も抽出され，この出現頻度の高い上位 5 つが，頻出ワードの候補として切り出される．

2.2 SOM について

SOM とは，人工ニューラルネットワークの一つで，データ間の類似性をマップ上の距離や位置関係として表現することができる教師なし学習モデルであり，データ間の空間的なつながりを低次元空間に写像することが可能である．例えば，簡単な例として，表 2 に示すような 3 次元の a,b,c の要素をもつデータ間の類似性を，SOM を用いてより低次元な 2 次元平面上に表現することを考える．

表 2：3 次元の要素をもつデータの例

データ	要素a	要素b	要素c
データ 1	0	0	0
データ 2	1	0	0
データ 3	1	1	1

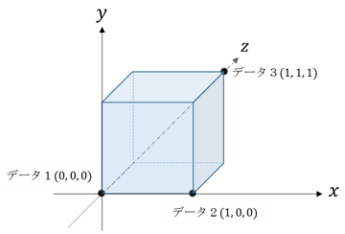

図 2：データの空間的な位置文章の記述

データの各要素の値を，それぞれの次元軸での座標としてみると，図 2 のようになる．
説明が簡単になるように，これを立方体の頂点座標として考えると，データ 1 (0,0,0) と データ 2 (1,0,0) は同じ辺にあり，近い位置関係にあるが，それに対してデータ 3 (1,1,1) は，2 つのデータから離れた位置にある．したがって，データ 3 は，他の 2 つのデータとは類似していないと考える．実際の空間的な距離は，ユークリッドの距離を用いて計算することで，データ間の類似性を数学的に評価できる．一般的な SOM では，データ間の類似性にもとづいた分類を行う過程で，このユークリッド距離が用いられる．

一般にn次元空間におけるユークリッド距離は，

$$E = \sqrt{\sum_{k=1}^{n}(u_k - v_k)^2}$$ で与えられる．

n 次元の要素をもつデータの類似度は，このようにして計算されるが，SOM におけるデータの類似度の計算には，ユークリッド距離を必ず用いなければならないものではなく，マハラノビス距離を用いた手法[4]や，モデル間の尤度の差を用いる手法などがあり，SOM は，ベクトルデータだけでなく様々なモデルの分類に応用ができる．本研究では，ウェブページの特徴を表すベクトルデータの類似性についての評価を検討するため，データ間の類似性にはユークリッド距離をベースとした評価方法を用いている．

SOM 具体的なアルゴリズムの説明については，2.3.2 の，「SOM を用いたウェブページの特徴の分類」の中で説明する．

2.3 性別・年代を考慮したウェブページの特徴の分類

アクセスされるウェブページの特徴は，性別・年代によって異なると考えられ，利用者の性別・年代も考慮したウェブページの分類が必要である．そこで，「Yahoo! JAPAN における 2015 年 1 月 1 日から 2015 年 12 月 31 日までの検索クエリ TOP30」のデータを用いた．これは，表 3 に示すように，性別・年代ごとの検索ワード上位 30 のデータであり，検索ワードは，その性別・年代が興味関心をもつ事柄として考えられる．しかし，検索ワードだけでは，どのようなウェブページに関心を持っているかというデータとしては不十分だと考えられる．そこで，検索ワードをブラウザで検索した際に，ウェブサイトの候補として表示される上位 5 つのサイトのトップページから html を取得し，2.1.1~2.1.2 の手法を用いて，その html におけるテキスト情

ⓒ2017 Information Processing Society of Japan

報から頻出ワードの上位 5 つを抽出する．抽出されたワードは，その検索クエリを検索したユーザが興味関心をもつウェブサイトの特徴として考える．このようにして，性別・年代ごとの TOP30 の検索クエリについてそれぞれ 5 つの頻出ワードを抽出し，すべての性別・年代で，6(年代)×2(性別)×30(検索クエリ)×5＝計 1800 個の頻出ワードを抽出する．このとき，各年代で同じ頻出ワードが抽出された場合でも，まとめることはせず，それぞれの抽出ワードとして考える．こうすることで，例えば 10 代の特徴でみられるような「ニコニコ動画」に関連したクエリが他の年代に比べて多いことから，10 代の検索クエリすべてから抽出される頻出ワード 300 個の中に，「ニコニコ動画」に関するワードが多く入ることになる．もし，同じ抽出ワードをまとめて一つとして考えた場合，ある年代である事柄に関連したクエリが多く検索されるという特徴を考慮することができなくなる．
これらの，1800 個の抽出ワードと，各性別・年代の情報を用いることで，各性別・年代を考慮したウェブページの特徴を表すベクトルデータを作成することを考える．
以下，具体的なベクトルデータへの変換方法について述べる．

2.3.1 性別・年代を考慮したウェブページの特徴のベクトル化について

Yahoo! JAPAN の公開している，検索クエリ全体から抽出した頻出ワード 1800 個と
性別・年代に関する要素をもつ図 3 に示すようなベクトルデータを考える．

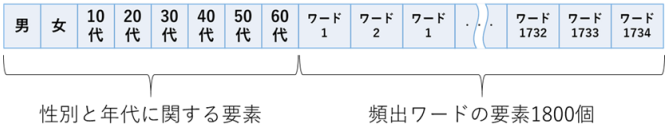

図 3：性別・年代を考慮したベクトルデータの要素

この要素に当てはまる項目には 1，当てはまらない場合には 0 の値をセットする．
例えば，10 代女性の検索クエリ「占いツクール」から得られた 5 つの頻出ワードが
ワード 1，ワード 3，ワード 42，ワード 55，ワード 1732 であったとすれば，1 をセットする要素の個数は，ワード 1 の要素が 2 箇所あることに注意して，合計 7 つである．

このようにして，それぞれの性別・年代における，各検索クエリのベクトルデータを
作成すると，360 個のベクトルデータが得られる．このようにして得られたベクトルデータを，SOM 入力データとして用いることで，性別・年代を考慮したウェブページの特徴を分類する．以下，具体的な分類方法について述べる．

2.3.2 SOM を用いたウェブページの特徴の分類

性別・年代ごとのベクトルデータを，SOM を用いて分類することを考える．
以下は，SOM のアルゴリズムの概要である．

性別・年代，頻出ワードの情報を表す 1808 個の要素をもった，360 個の入力データを

$$\boldsymbol{v_n} = (v_{n,1}, v_{n,2}, v_{n,3}, \dots, v_{n,1808}) \quad n = 1,2,3, \dots,360$$

とする．ここで，前節 2.3.1 で示したように，v_1は男性の要素，v_2は女性の要素，v_3~v_8は年代に関する要素，v_9~v_{1808}は頻出ワードに関する要素である．また，マップ上のノードにおけるベクトルを

$$\boldsymbol{m_{x,y}} = (m_{x,y,1}, m_{x,y,2}, m_{x,y,3}, \dots, m_{x,y,1808})$$

とする．本研究で用いた SOM のマップサイズは 30×30=900 であり，x, y は
$0 \le x < 30, 0 \le y < 30$ を満たす整数である．

STEP 1：
マップをランダムな値で初期化する

マップ上のベクトル$\boldsymbol{m_{x,y}}$の値を，乱数で初期化する．

STEP 2：
入力データ$\boldsymbol{v_n}$ に最も類似したベクトルをマップ上から探す

類似度は入力データ$\boldsymbol{v_n}$と，マップ上のベクトル$\boldsymbol{m_{x,y}}$とのユークリッド距離で与えられ，

$$\min_{x,y} \sqrt{\sum_{k=1}^{1808} (v_{n,k} - m_{x,y,k})^2}$$

を満たすx, y が入力データに最も類似したマップ上の位置となる．
この位置を(x^*, y^*)とする．これが勝者ノードといわれる位置であり，それぞれの入力データ$\boldsymbol{v_n}$は，それぞれの勝者ノードの位置に写像される．

STEP 3：
勝者ノードの近傍にあるマップ上のベクトルを，入力データ$\boldsymbol{v_n}$に近づける

勝者ノードとマップ上のノードとの距離についても，ユークリッド距離で求められ，
$\sqrt{(x^* - x)^2 + (y^* - y)^2}$ で計算できる．勝者ノードからの近傍の距離を$f(t)$とすれば，

ⓒ2017 Information Processing Society of Japan

マップ上にある近傍のベクトルの位置x, yは次の式を満たす.

$$\sqrt{(x^* - x)^2 + (y^* - y)^2} < f(t)$$

ただし，$f(t)$は学習回数 t とともに，減少する関数でなければならない.
この式を満たす，マップ上のベクトルを$\boldsymbol{m}_{x,y}^*$とすると，次のように, 勝者ノードの近傍にあるベクトル$\boldsymbol{m}_{x,y}^*$ は更新される.

$$\boldsymbol{m}_{x,y} = \boldsymbol{m}_{x,y}^* + \boldsymbol{\alpha}(t, f(t)) \times (\boldsymbol{v}_n - \boldsymbol{m}_{x,y}^*)$$

ここで，$\boldsymbol{\alpha}(t, f(t))$は学習率とよばれ，学習回数$t$，近傍距離$f(t)$が大きくなれば，減少するような関数を用いる．本研究では，$f(t)$は直線の関数で，学習回数$t = 1$回目，$t = 100$回目での値はそれぞれ,

$$f(1) = \frac{15}{2}\ ,\quad f(100) = 1.5$$

となるように設定しており,

$$f(t) = -\frac{2}{33}(t-1) + \frac{15}{2}$$ である.

STEP 4：
学習回数分　繰り返し学習を行う

STEP 2~STEP3 の操作を学習回数分ループさせて，マップを更新させていく.
入力データの最終的な勝者ノードの位置関係が，それぞれの入力データの類似関係となる.

このステップで最終的にできるマップは，ウェブページの特徴を類似度に応じて分類したようなマップとなる．したがって，あるマップ上の位置にある $\boldsymbol{m}_{x,y}$ の値を参照することで，その位置に分類されたウェブページの特徴を知ることができる.

このようにして得られたマップを用いて，実際にインターネット利用者の性別・年代の推定方法について述べる.

2.4 SOM を用いたインターネット利用者の性別・年代の推定実験

推定に用いるデータは，10 代から 60 代までの被験者の利用するウェブページの各 10 個のデータを用意した．しかし，パソコンを使用する機会が少ないなどの理由で，10 代女性，60 代女性のデータは取得できなかった.

推定の実験方法は，前節 2.3.2 の手法であらかじめウェブページの特徴を，SOM を用いてマッピングしてく．次に被験者の利用したウェブページが，マップ上にあるウェブページのどの特徴に類似しているかを求め, どのような性別・年代から興味関心をもたれやすいウェブページであるかを推定することを考える.

具体的な手法を以下に示す.

STEP 1：被験者の利用するウェブページに含まれる，上位５つの頻出ワードを前節 2.1.1~2.1.2 の手法を用いて抽出し，これを，前節 2.3.1 の手法を用いてベクトルデータに変換する.

STEP 2：各被験者におけるベクトルに最も類似したベクトルをマップ上から探す．つまり，各被験者の勝者ノードをマップ上から探す.
この勝者ノードにあるウェブページの特徴は，被験者が利用するウェブページの特徴に類似していると考えられる.

STEP 3：各被験者における勝者ノードの位置にあるベクトルの性別・年代に関する要素を参照し，これをもとに被験者の性別・年代と推定する．年代の推定方法は要素の 10 代を 15 歳，20 代を 25 歳，30 代を 35 歳，以下同様その年代の中央値を各年代の代表値として考え，各年齢の中央値を変数 XとしてX $= \{x_1, x_2, x_3, x_4, x_5, x_6 |\ x_t = 10t + 5\}$ と定義する．また勝者ノードの位置にあるベクトル 10～60 代の要素の値をこの年代順に$\{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$とすると，被験者の年齢がX $= x_t$ である確率P(X $= x_t$)は，P(X $= x_t) = \omega_t$ で与えられるから, 被験者の年齢は変数 X の期待値で定義でき,

$$\sum_t x_t\, P(X = x_t) = \sum_t x_t\, \omega_t$$

であるから, 被験者の年齢は　$x_t\omega_t$ の総和をとることで推定できる.

被験者の性別については，勝者ノードの位置にあるベクトルの性別に関する要素うち,
大きい値をもつ性別を, その被験者の性別として推定する.

このような方法で推定した被験者の性別・年代に関する実験結果を次に示す.

ⓒ2017 Information Processing Society of Japan

2.5 被験者の性別・年代に関する推定結果

被験者のデータは，アクセスしたウェブページ１０個分のデータで，各ウェブページにおいて推定される利用者の性別・年代をそれぞれ算出し，それらの平均値を取ったものを推定値として考えた．

それぞれの性別・年代ごとに推定された結果を表 3，表 4 に示す．

表 3：性別・年齢の推定結果

性別	年代	推定された性別	推定値（年齢）
男性	１０代	女性	１５歳
男性	２０代	女性	３６歳
男性	３０代	男性	３２歳
男性	４０代	男性	４０歳
男性	５０代	女性	５１歳
男性	６０代	男性	２６歳
女性	２０代	女性	１５歳
女性	３０代	男性	５６歳
女性	４０代	女性	３５歳
女性	５０代	女性	３５歳

表 4：性別における推定精度

性別	精度
男性	５０％
女性	７５％
男女トータル	６０％

男性が誤って女性と推定される割合は 50%となり，性別の推定が適切にできていないことが分かる．
また，女性が男性と誤って推定されたのは 30 代女性のみであった．

性別においては，男性のほうが女性に比べて推定できていない結果となった．一方，年代における推定では，男性は約 67%の精度で年齢を推定できているが，女性は年齢を推定できているとはいえない結果となった．
この 67%という数値は，年代が 10 代から 60 代の 6 パターンについて，6 データ中 4 つ偶然当たる確率が約 0.8%であることを考えると十分有用な結果といえる．

3. おわりに

本研究では，ＳＯＭを用いた分類のアルゴリズムだけでなくウェブページから頻出ワードを抽出するアルゴリズムについても開発を行った．実際の性別・年代の推定結果は，男性では性別が，女性では年代の推定が適切にできていなかった．原因としては，マップを作成する際に，タグなどの意味のない情報までマップの学習に使用されていたことや，「yahoo」や「ブログ」，「All」，などのよく出現するワードが女性的なワードとして広くマップ上に分類されていたことなどが原因として考えられる．

また，被験者がある一つのサイト内だけのウェブページを多く申告したデータがあることによって，一部の偏った特徴に結果が引きずられてしまうなど，被験者のデータの収集にも課題が残る．

今回の実験では，性別・年代の推定に関する有用な結果を得ることができたが，頻出ワードを抽出するアルゴリズムや，ウェブページからテキストを適切に抽出するアルゴリズムにおいてまだ改善すべき点も多くあることも分かった．今後は，これらのアルゴリズムを改善することで，より適切な推定を目指すとともに，被験者の適切なデータ収集についても検討する．

参考文献

[1] 坪坂 正志, “言語モデルを使ったクエリログからの性別推定,”情報処理学会研究報告, Vol.2013-SLP-96, No.14, 2013.
[2] T. Kohonen, “The Self-Organizing Map,” Proceedings of the IEEE, 78, 1464-1480, 1990.
[3] T. Kohonen, “Self-Organized Formation of Topologically Correct Feature Maps,” Biological Cybernetics, Springer-Verlag 1982.
[4] Sourav Paul , Mousumi Gupta, “Image Segmentation By Self Organizing Map With Mahalanobis Distance”, International Journal of Emerging Technology and Advanced Engineering, (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 2, February 2013.

ⓒ2017 Information Processing Society of Japan