スマートフォンベースユーザ参加型主観的情報データ分析システム

市村　匠

スマートフォンベースユーザ参加型システムで収集された主観的データをソフトコンピューティングの手法を用いて分析し，特徴を抽出する方法がある．本論文では，Android スマートフォンアプリケーションとして開発された，「広島県観光マップ」により収集された情報を分析し，新しい観光地の特徴を抽出する．ここでは，インタラクティブ階層成長型自己組織化マップによる分類結果をもとに，C4.5 でフィルタリングルールを抽出した結果を報告する．

キーワード：Android スマートフォン，スマートフォンベースユーザ参加情報収集システム，ソフトコンピューティング

1. はじめに

今日，クラウド・コンピューティングなど，情報技術の進歩により，実世界に存在する膨大で，かつ不確実性要素を含むデータの収集が容易になっている．なかでも携帯電話を用いたデータ収集の方法は，‘Mobile Phone based Participatory Sensing (MPPS)’[2] として知られている．この手法は，ユーザの位置情報やセンサーで感知できる情報を，携帯電話を通じて自動で収集する．スマートフォンと呼ばれる多機能携帯電話では，GPS，運動センサー，カメラなどのセンサーがあり，携帯電話回線だけでなく WiFi などの無線 LAN 通信を利用することができる．このような情報収集の手段は，観測機器を設置して測定する「定点観測」より，参加するユーザの数だけ多くの観測点を設けられるため，台風や震災時のリアルタイム情報収集手段として注目されている．

スマートフォンの一つとして知られている Android は，オープンソースで提供される開発者キットを用いれば，自由にアプリケーションを開発することができる．例えば，撮影した写真に GPS により測定した位置情報を埋め込むような機能や，コメント文などの自然言語を入力すると同時に含まれている単語を抽出する機能など，開発者が独自の機能を開発できる．

われわれは，Android スマートフォンを用いた広島県観光マップ [1] を開発し，Google Play で無料公開している．このアプリケーションは，広島県の観光情報を提供するだけでなく，GPS による位置情報，位置情報から自動変換された住所，5 段階による評価，カメラで撮影した写真，自由記述によるコメント文を収集することができ，観光客が実際に観光地に訪れたときに，観光客が感じた情報を投稿できるシステムになっている．このように収集されたデータには，観光客の主観的情報が含まれていると考えられ，分析結果により広島県観光の魅力あるスポットを発見することができ，観光産業を促進するための知識獲得が期待されている．

収集された主観的情報には，欠損値や矛盾だけでなく，曖昧な情報が含まれている場合がある．このようなデータの分析には，厳密な解を求めるより，柔軟な問題解決の手法として知られるソフトコンピューティングを用いるほうが効果的であることが多い．ソフトコンピューティングには，ニューラルネットワーク，進化計算手法などさまざまな方法が知られているが，本論文では，主観的情報の分析手法として，インタラクティブ GHSOM (Growing Hierarchical SOM)[5] の手法による分類結果から C4.5 を用いて知識獲得を行う手法について説明する．

GHSOM は階層成長型自己組織化マップ (Growing Hierarchical SOM)[4] であり，自己組織化マップ (SOM)[3] がベースとなった教師なし競合学習アルゴリズムとして知られており，階層構造を持った複数の独立した SOM から構成される（図 1）．SOM は入力データを 2 次元空間に写像できるため，入力データの特徴を視覚的にとらえやすいアルゴリズムである．学

いちむら　たくみ
県立広島大学 経営情報学部 経営情報学科
〒 734–8558 広島市南区宇品東一丁目 1 番 71 号

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

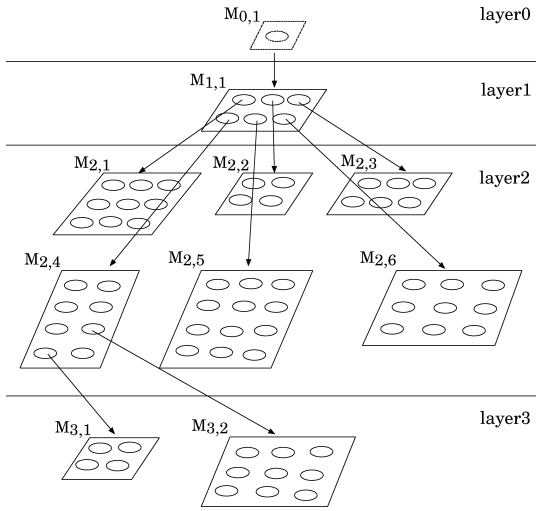


図 1　GHSOM の階層構造

習の過程でユニット挿入と階層化を行うことで，入力データに適応するように，マップサイズが動的に成長する．しかし，パラメタの設定によっては深く階層化されすぎ，かえって入力データの特徴がとらえづらくなる場合がある．この問題に対して，階層化を抑制する手法が，インタラクティブ GHSOM[5] である．この方法は，階層化し過ぎた場合でもマップを再構築し，複雑になってしまった構造を修正するもので，計算結果から部分的に階層構造を変更することができる．

本論文では，スマートフォンベースユーザ参加型主観的情報データ分析システムである，広島県観光マップで収集された情報をインタラクティブ GHSOM と C4.5 によって，得られた広島県観光に関する特徴を抽出した結果を説明する [6, 7].

2. インタラクティブ GHSOM

本節では，まず GHSOM のアルゴリズムを説明し，パラメタ設定に関する難しさを軽減するために，人間の感覚に近いパラメタの与え方について説明する．

2.1 GHSOM のアルゴリズム

Step1)　GHSOM のすべてのパラメタを初期化する．

Step2)　層 0 の平均量子化誤差を計算する．

$$mqe_0 = \frac{1}{n_I}\sum_{\mathbf{x}_i \in I} \| \mathbf{m}_0 - \mathbf{x}_i \|, n_I = |I| \quad (1)$$

ここで $\mathbf{m}_0$ は入力データの平均，n_I は入力データの数，I は入力データ集合，$\mathbf{x}_i$ は入力ベクトルである．

Step3)　層 ℓ ($\ell = 1, 2, \ldots$) ごとにラベル付けしたマップ $\mathbf{M}_{\ell,v}(v = 1, 2, \ldots)$ をマップサイズ ($2 \times 2 (= u_0)$) とし，SOM により入力データを分類する．ここで k を勝者ユニットとし，分割された勝者の入力ベクトル集合を I とする．

Step4)　層 ℓ での同じマップ $\mathbf{M}_{\ell,v}$ における各々の勝者ユニットに対する量子化誤差とその平均を計算する．

平均量子化誤差：

$$mqe_k = \frac{1}{n_C}\sum_{\mathbf{x}_j \in \mathbf{C_k}} \| \mathbf{m}_k - \mathbf{x}_j \|, \quad (2)$$
$$n_C = |\mathbf{C}_k|, \mathbf{C}_k \neq \phi$$

量子化誤差：

$$qe_k = \sum_{\mathbf{x}_j \in \mathbf{C_k}} \| \mathbf{m}_k - \mathbf{x}_j \|, \quad (3)$$
$$n_C = |\mathbf{C}_k|, \mathbf{C}_k \neq \phi$$

ここで $\mathbf{m}_k$ は勝者ユニット k の参照ベクトル，$\mathbf{C}_k$ は勝者ユニット k に割り当てられた入力ベクトル集合，$\mathbf{x}_j$ は集合 $\mathbf{C}_k$ の要素である．

Step5)　層 ℓ の同じマップ $\mathbf{M}_{\ell,v}$ における勝者ユニットの参照ベクトル中，入力ベクトルと最も大きな量子化誤差を持つ参照ベクトルに接続されるユニットを誤差ユニット e として選ぶ．誤差ユニット e とその近傍ユニットの参照ベクトルを比較し，差が最も大きい近傍ユニット d を選ぶ．

Step6)　層 ℓ の同じマップ $\mathbf{M}_{\ell,v}$ における勝者ユニット k の部分集合 $\mathbf{u}$ の量子化誤差 qe_k の平均 $mqe_{\mathbf{M}_{\ell,v}}$ を計算する．

$$mqe_{\mathbf{M}_{\ell,v}} = \frac{1}{n_u}\sum_{k \in \mathbf{u}} qe_k, n_u = |\mathbf{u}| \quad (4)$$

Step7)　式 (5) を満たすとき，図 2 に示すように，ユニットの行または列を誤差ユニット e とその最も異なる近傍ユニット d の間に挿入する．

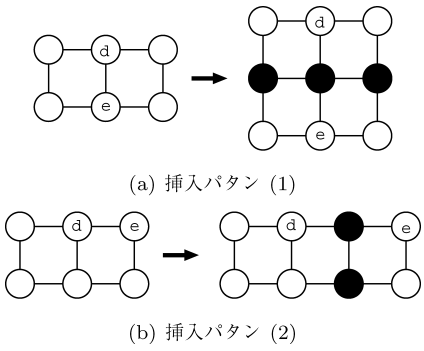
$$mqe_{\mathbf{M}_{\ell,v}} \geq \tau_1 qe_w \quad (5)$$

ここで，w は層 $\ell - 1$ のマップを構成しているユニットを示す．

Step8)　新たに挿入された参照ベクトルの初期値は誤差ユニット e とその最も異なる近傍ユニット d の間にそれらのユニットの重みの平均として与える．

Step9)　式 (5) を満たさなくなると，ユニットは挿入されなくなる．ユニットが挿入されなくなると，上位層まで戻り，ほかのユニットから生成された

 Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.



(a) 挿入パタン (1)

(b) 挿入パタン (2)

図 2 ユニットの挿入

マップについて，ユニット挿入のアルゴリズムを適用する．

Step10) ユニットの挿入過程が終了した後，階層化を行うかどうかの判定を行う．式 (6) を満たしている場合に階層化を行う．

$$qe_k \geq \tau_2 qe_0 \tag{6}$$

2.2 GHSOM の成長制限のパラメタ

GHSOM は図 1 で示すように，階層構造を持った複数の独立した SOM から構成される．提案手法のインタラクティブ GHSOM では，1 つ上の層にユニットを挿入することで階層化を抑制する方法を提案した．

Case 1) 階層化の条件を満たし，かつ式 (7) を満たす場合，つまり，ある勝者ユニットに対応する入力データのサンプル数が全体の入力データのサンプル数と比べてかなり少ないとき，階層化を停止し，再度ユニットの挿入を行う．ただし，量子化誤差が予め定められた値より小さい場合，ユニット挿入は行わない．

$$n_k \leq \alpha n_I \tag{7}$$

ここで，n_k と n_I はそれぞれ勝者ユニット k に対応する入力データの数，全入力データ I の数であり，α はパラメタである．

Case 2) 階層化が行われないとき，かつユニット全体量子化誤差が大きいとき，そのユニットの量子化誤差が同じ層のユニットの量子化誤差の合計に対し大きいなら，式 (8) を満たすと再度ユニットの挿入を行う．

$$qe_k \geq \beta\tau_1 \sum qe_y, \quad y \in \mathbf{S}_k \tag{8}$$

ここで，$\mathbf{S}_k$ は 1 つ上の層の勝者ユニット k の集合である．

3. 広島県観光マップ

図 3(a) は，開発した広島県観光マップの初期画面である．実際に訪れた観光地では，図 3(b) を用いて，GPS による位置情報，5 段階評価値，140 文字以内のコメント文，写真を収集すると同時に，サーバにデータを投稿することができる．これらのデータをほかのスマートフォンでも利用でき，図 4 のように，観光情報を KML フォーマットに変換し，パソコンでも GoogleMap 上に表示できるクラウドサービスを提供している．

4. GHSOM による分類と C4.5 による知識獲得

4.1 観光 Web サイトからの TF–IDF

広島県観光マップには，多値である評価，数値情報である位置情報，自然言語から構成されるコメント，観光地名がある．このうち，コメントは短文であるために，含まれている情報量が多くない．このため，広島県観光ホームページから情報を補完することを目的として，表 1 から TF–IDF 値 [8] を求め，コメント文とあわせて分類した．

4.2 GHSOM による分類

本節では，従来型の GHSOM と提案した GHSOM を用いて，観光マップで収集された 500 件の主観的データを分類した．図 5(a) は提案した GHSOM による分類結果を示している．また，図 5(b) は提案した GHSOM による分類結果をもとに，ユーザが特定のユニットを選択し，その部分だけ再学習した結果を示している．ここで用いたパラメタは，$\tau_1 = 0.1$, $\tau_2 = 0.01$, $\alpha = 0.03$, $\beta = 0.6$ であった．

分類結果を示すインターフェースでは，単純に得られた構造を画面上に描画するだけでなく，生じたマップ構造の任意の部分を選択し，その部分のみを再学習し，構造を再構築することができる．ここで，類似されたサンプルは同じノードに分類されるが，この状態をわかりやすく表示するために，マンセルカラーシステム [9] によって，式 (9) によって，RGB から変換した値によって，その類似度を円環状に示している（図 6）．

$$\tan(h_{rgb}) = \frac{\sqrt{3}(G-B)}{2R-G-B} \tag{9}$$

ここで，R, G, B は赤，緑，青の三原色の数値を示している．分類された結果，葉ノードのマップにあるユニットには，サンプルが割り当てられている．どのようなサンプルかをそのまま閲覧できるように，該当す

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

(a) 初期画面

(b) 新規投稿画面

図 3　広島県観光マップ

るユニットをクリックすると，そこに分類されているサンプルが表示される．また，分類されたサンプルが極端に少ないマップであれば，逆にどのような分類が行われたのか，人間にとっては細かすぎて理解できないことがある．人間の主間的評価をもとに満足する解が得られるまで解を探索する，インタラクティブな手法を実現している．これは，冗長な分類を行ったノードをクリックすることで，その部分だけ再学習を行い，粗い分類を行うことで，人間には直感的に理解しやすい分類結果を与えている．

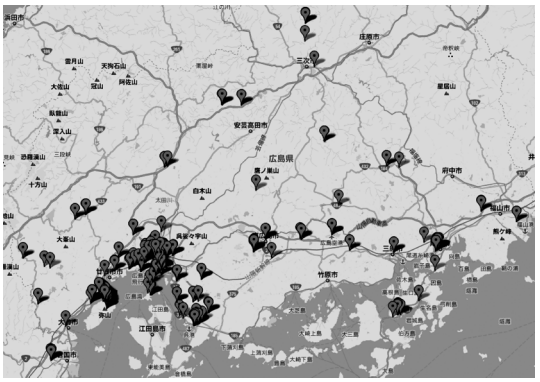


図 4　GoogleMap 表示

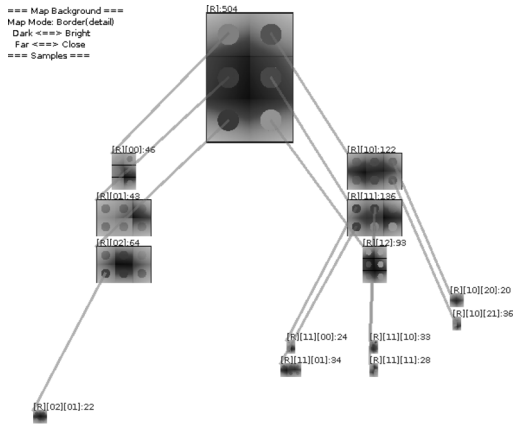
表 1　広島県の観光 Web サイト

観光協会による Web サイト			
Web サイト	サイズ (KB)	文書数	単語数
広島県観光ホームページ	3,100	1,368	11,982
呉観光協会	50	12	1,247
はつかいち旅ナビ	113	134	1,552
おのなび	1,100	410	5,600
福山観光協会	692	693	3,449
三次市観光協会	115	41	879
安芸高田市観光情報	248	227	2,044
観光客によるブログ			
Web サイト	サイズ (KB)	文書数	単語数
なっちゃんの遊び場ネット	512	160	3,391
広島おでかけマップ	979	372	5,290

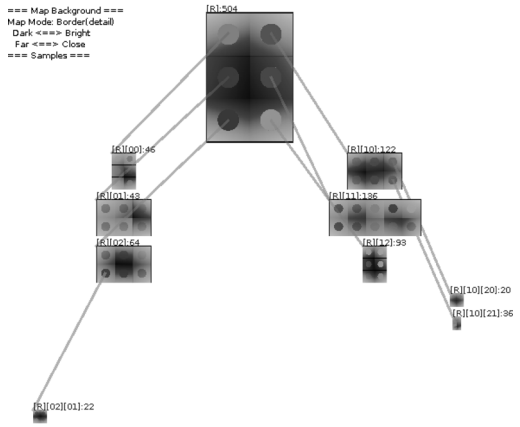
4.3　C4.5 による知識獲得

教師データ集合から決定木を生成する手法である決定木学習の代表的な手法として，C4.5 が知られている [10]．観光マップのサンプルに対し，GHSOM で分類された結果をクラスとして与え，C4.5 による決定木を作成した．図 5 では教師なし学習であるインタラクティブ GHSOM による分類結果が得られた．この結果を出力信号として考え，収集したデータとあわせることで，教師あり学習法である C4.5 に適用することが可能となる．図 7 は，その一部であるが，提案した GHSOM 手法による分類結果をもとに C4.5 の解析をしたもの（図 7(a)）とインタラクティブ GHSOM 手法による分類結果をもとに C4.5 の解析をしたもの（図 7(b)）を比較した．図中のアルファベットは GH-

 Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.



(a) GHSOM



(b) インタラクティブ GHSOM

図 5　GHSOM による分類結果

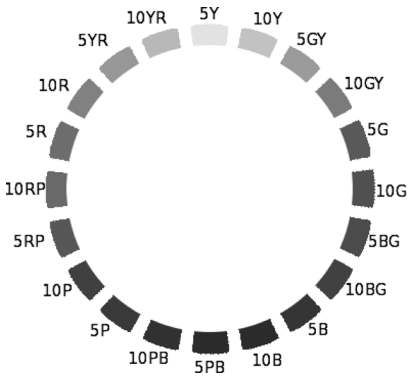


図 6　マンセルカラーシステム

SOM の葉ノードを示しており，再学習前では 9 つのクラスタがあったが，再学習後には，4 つのクラスタに再分類された．これらの違いから得られた知識の再編成が行われていることがわかる．

```
evaluation > 2 :
|  evaluation <= 3 :
|  |  tfidf <= 1.01692 :
|  |  |  lon <= 132.485 :
|  |  |  |  lon <= 132.328 :
|  |  |  |  |  tfidf <= 0.4022 : [R][11][11][01] (8.0)                    A1
|  |  |  |  |  tfidf > 0.4022 :
|  |  |  |  |  |  lat <= 34.4323 : [R][11][21] (14.0/1.0)                 B
|  |  |  |  |  |  lat > 34.4323 : [R][11][20] (2.0)                       C
|  |  |  |  lon > 132.328 :
|  |  |  |  |  tfidf <= 0.536494 :
|  |  |  |  |  |  tfidf > 0.329448 : [R][11][11][12] (8.0/1.0)            E1
|  |  |  |  |  |  tfidf <= 0.329448 :
|  |  |  |  |  |  |  lat <= 34.3906 : [R][11][11][00] (2.0)               A2
|  |  |  |  |  |  |  lat > 34.3906 : [R][11][11][11] (8.0/1.0)            E2
|  |  |  |  |  tfidf > 0.536494 :
|  |  |  |  |  |  tfidf <= 0.72588 : [R][11][10][00] (6.0)                F1
|  |  |  |  |  |  tfidf > 0.72588 :
|  |  |  |  |  |  |  lat <= 34.33 : [R][11][10][11] (2.0)                 F2
|  |  |  |  |  |  |  lat > 34.33 : [R][11][10][10] (15.0)                 F3
|  |  |  lon > 132.485 :
|  |  |  |  lon <= 132.63 :
|  |  |  |  |  tfidf > 0.650069 : [R][11][10][12] (6.0/1.0)               D
|  |  |  |  |  tfidf <= 0.650069 :
|  |  |  |  |  |  tfidf <= 0.419397 : [R][11][11][02] (3.0)               E3
|  |  |  |  |  |  tfidf > 0.419397 : [R][11][10][01] (4.0/1.0)            E4
|  |  |  |  lon > 132.63 :
|  |  |  |  |  lon <= 132.967 : [R][01][21] (7.0/2.0)
|  |  |  |  |  lon > 132.967 : [R][01][00] (10.0)
```

(a) GHSOM 手法

```
evaluation > 2 :
|  evaluation <= 3 :
|  |  tfidf <= 1.01692 :
|  |  |  lon <= 132.386 :
|  |  |  |  tfidf <= 0.4022 : [R][11][21] (10.0)                          A'
|  |  |  |  tfidf > 0.4022 :
|  |  |  |  |  lat <= 34.4323 : [R][11][40] (15.0/2.0)                    B
|  |  |  |  |  lat > 34.4323 : [R][11][31] (2.0)                          C
|  |  |  lon > 132.386 :
|  |  |  |  lon <= 132.63 :
|  |  |  |  |  tfidf <= 0.603696 :
|  |  |  |  |  |  lat > 34.4094 : [R][11][31] (3.0/1.0)                   C
|  |  |  |  |  |  lat <= 34.4094 :
|  |  |  |  |  |  |  lat <= 34.2422 : [R][11][41] (2.0/1.0)               D2
|  |  |  |  |  |  |  lat > 34.2422 : [R][11][20] (20.0)                   E'
|  |  |  |  |  tfidf > 0.603696 :
|  |  |  |  |  |  lat <= 34.2704 : [R][11][41] (5.0)                      D1
|  |  |  |  |  |  lat > 34.2704 : [R][11][30] (21.0)                      F'
|  |  |  |  lon > 132.63 :
|  |  |  |  |  lon <= 132.967 : [R][01][21] (7.0/2.0)
|  |  |  |  |  lon > 132.967 : [R][01][00] (10.0)
```

(b) インタラクティブ GHSOM 手法

図 7　C4.5 による獲得された知識

4.4　抽出されたルール

C4.5 で得られたルールはいずれも，位置情報，TF-IDF，評価値の入力項目を前件部として，インタラクティブ GHSOM によって得られたクラスを後件部とした形式で獲得される．一般的な手法による解析結果では，データ数が多いものがルールとして抽出される．このことは，情報量が少ないために，希少ではあるが重要だと考えられるルールが抽出されない場合があることを意味している．本手法では，このようなルールでも抽出できるように再分類をインタラクティブに行

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

うことが可能となっている．

本実験において，再分類の結果，C4.5 により生成された木の構造は，図 7(a) は図 7(b) に変化したが，この結果から，次のようなことがわかる．つまり，A_1，A_2 が A' に，E_1，E_2，E_3，E_4 が E' に，F_1，F_2，F_3 が F' にそれぞれ統合されている．一方，D は新しく D_1 と D_2 に分割されている．図 8 は，A，B，D に該当するサンプルのみを撮影した写真で表現することで抽出された知識をわかりやすく表現したものである（E，F は観光地の違いだけで同様な知識が抽出されていた）．

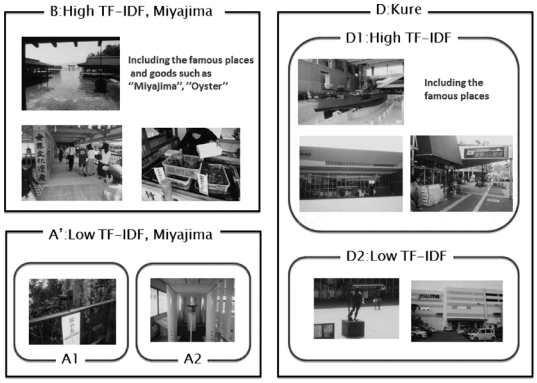


図 8　獲得されたフィルタリングルール

5. まとめ

MPPS として広島県観光マップアプリを開発し，ユーザが実際に観光した場所で，個人の主観的情報を投稿できる Android アプリケーションを開発し，公開している．収集された情報を分析するために，人間の直感的な感覚に基づいて GHSOM のパラメタを設定する手法を提案し，さらに得られた分類結果において，人間がわかりづらい箇所があれば，その部分だけ再学習するようなインタフェースを開発した．提案した手法によって，観光情報を分析した結果，観光地や土産物として有名でなくても，ユーザから良い評価が与えられる情報があれば，新しく発見された観光名所，特産品を Twitter で発信するためのフィルタリングルールを抽出した．今後は，MPPS を通じて発見された情報を交換するために，CGM (Consumer Generated Media) として，広島県観光コミュニティを構築する予定である．

参考文献

[1] ITProducts，広島観光マップ，https://play.google.com/store/apps/details?id=jp.itproducts.KankouMap, Retrieved 2012-12-17.

[2] N. D. Lane, E. Miluzzo, L. Hong, D. Peebles, T. Choudhury, and A. T. Campbell, A survey of mobile phone sensing, *IEEE Communications Magazine*, **48**(9), 140–150, 2010.

[3] T. Kohonen, *Self-Organizing Maps*, Springer Series in Information Sciences, **30**, Springer, Berlin, Heidelberg, New York, 1995.

[4] A. Rauber, D. Merkl, and M. Dittenbach, The growing hierarchical self-organizing map: Exploratory analysis of high-dimensional data, *IEEE Transactions on Neural Networks*, **13**(6), 1331–1341, 2002.

[5] T. Ichimura and T. Yamaguchi, A Proposal of Interactive Growing Hierarchical SOM, *IEEE International Conference on Systems, Man, and Cybernetics*, 3149–3154, 2011.

[6] T. Ichimura, S. Kamada, and K. Kato, Analysis of Tourist Subjective Data in Smartphone based Participatory Sensing System by Interactive Growing Hierarchical SOM, *Proc. of 4th International Conference on Intelligent Decision Technologies (KES IDT 2012)*, 764–773, 2012.

[7] T. Ichimura, S. Kamada, and K. Kato, Knowledge Discovery of Tourist Subjective Data in Smartphone Based Participatory Sensing System by Interactive Growing Hierarchical SOM and C4.5, *Intl. J. Knowledge and Web Intelligence*, **3**(2), 110–129, 2012.

[8] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok, Interpreting TF-IDF term weights as making relevance decisions, *ACM Transactions on Information Systems*, **26**(3), 137, 2008.

[9] R. G. Kuehni, The early development of the Munsell system, *Color Research and Application*, **27**(1), 2027, 2002.

[10] J. R. Quinlan, Improved use of continuous attributes in C4.5, *Journal of Artificial Intelligence Research*, **4**(1), 77–90, 1996.

 Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.