Available online at www.sciencedirect.com

SCIENCE DIRECT®

Fuzzy Sets and Systems 148 (2004) 43–60

FUZZY sets and systems

www.elsevier.com/locate/fss

Fuzzy integration of structure adaptive SOMs for web content mining

Kyung-Joong Kim, Sung-Bae Cho*

Department of Computer Science, Yonsei University, 134 Shinchon-dong, Sudaemoon-ku, Seoul 120-749, Republic of Korea

Abstract

Since exponentially growing web contains giga-bytes of web documents, users are faced with difficulty to find an appropriate web site. Using profile, information retrieval system can personalize browsing of the web by recommending suitable web sites. User's evaluation on web content can be used to predict users' preference on web sites and construct profiles automatically. User profile represents different aspects of user's characteristics, thereby we need an ensemble of classifiers that estimate user's preference using web content labeled by user as "like" or "dislike." Fuzzy integral is a combination scheme that uses subjectively defined relevance of classifiers and structure adaptive self-organizing map (SASOM) is a variant of SOM that is useful to pattern recognition and visualization. In this paper, fuzzy integral-based ensemble of SASOMs trained independently is used to estimate user profile and tested on UCI Syskill & Webert data. Experimental results show that the proposed method can perform better than not only previous naïve Bayes classifier but also majority voting of SASOMs.

Keywords: User profile; Web content mining; Structure adaptive self-organizing map; Fuzzy integral; Syskill & Webert

1. Introduction

Web is a new source of information but has no central control to maintain regularity like database. Because anyone can create web pages and link each other, web contains significantly huge number of documents. Nobody can use this information fully without knowledge discovery or advanced information search methods such as web search engine, web directory, and web mining. Web mining can be classified into three components according to the sources of information to discover knowledge: web content mining, web usage mining and structure mining. Web content mining analyzes HTML

* Corresponding author. Tel.: +82-2-2123-2720; fax: +82-2-365-2579.
E-mail address: sbcho@csai.yonsei.ac.kr (S.-B. Cho).

0165-0114/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.fss.2004.03.005

documents to provide the user with useful information by classifying web documents or creating profile to present user's preference based on the terms in the documents [18,28]. Web usage mining uses web log file, click stream, purchase record, and query log to extract useful knowledge [32]. Web structure mining exploits the link structure of web sites to extract relevant web pages or detect cyber community that has strong connectivity [3].

In this paper, we focus on web content mining for creating user profile from the HTML documents and user's preference record. Estimating user profile needs non-linear function because it has the properties that are not easily captured by simple guess. Also, it contains many aspects of user's preference such as "I like a web page that has a funny story," "I like a web page that is likely related to sports," and "I like a web page that has easily understandable words." It is difficult to estimate such properties with a single machine learning model and we need to combine a number of models that complement each other with different expertise.

In the meantime, self-organizing map (SOM) is a very useful neural network to visualize large-dimensional data for mining knowledge and an efficient tool to cluster data [13,37]. Some researchers attempt to apply SOM to pattern classification [4,34]. Like other models of neural networks, one of the shortcomings is the difficulty to determine the size and structure of the network. In the previous work [5] we proposed an efficient pattern recognizer based on a dynamic node splitting scheme for the SOM, which shapes the structure-adaptive SOM (SASOM) into a pattern recognizer by splitting a node representing more than one class into a submap (composed of four nodes). Ensemble of SASOM trained independently using different feature sets provided high performance in digit recognition problem [6].

SASOM can be used as an effective pattern recognizer and also visualize map structure to understand 2-D projection of user input vector. In this paper, we have adopted the ensemble of SASOM's to estimate user profile and each SASOM is trained independently using different feature sets. Three different feature ranking methods are used for this problem: information gain, TFIDF, and odds ratio. Information gain is a very effective feature extraction method that is used to induce decision tree in C4.5 [30]. TFIDF is a general method that is frequently used in text retrieval. Odds ratio is a very simple method that ranks features [25]. These three methods are representative feature ranking methods for text and simply implemented. There are many combination methods such as Borda count, Condorect function, voting, weighted voting, Bayesian averaging, Dempster-Shafer theory, and behavior knowledge space [36]. However, these methods cannot incorporate user's subjective preference on classifier into combination procedure and have little flexibility. Fuzzy integral is a combination method to aggregate evidence from multiple sources using fuzzy measure and user's subjective evaluation on classifiers' relevance [7]. In this paper, we propose fuzzy integral-based ensemble of SASOM's to estimate user profile from HTML documents labeled by user as "like" or "dislike."

Fig. 1 shows the overview of the proposed scheme. Web text data is a source of training and test and must be preprocessed to be used for input data. From a preprocessed feature vector for each web document, each feature extraction method selects relevant feature sets for training: each feature set is used to train one SASOM. After training, each SASOM has different topology as shown in this figure. Fuzzy integral aggregates evidence from multiple sources at the end. This ensemble classifier can be used to predict user's preference on unknown web documents as a user profile. To evaluate the proposed method, UCI KDD Syskill & Webert data is adopted [11]. The data contains four different topics and related web pages with user's preference record that is "hot," "medium,"

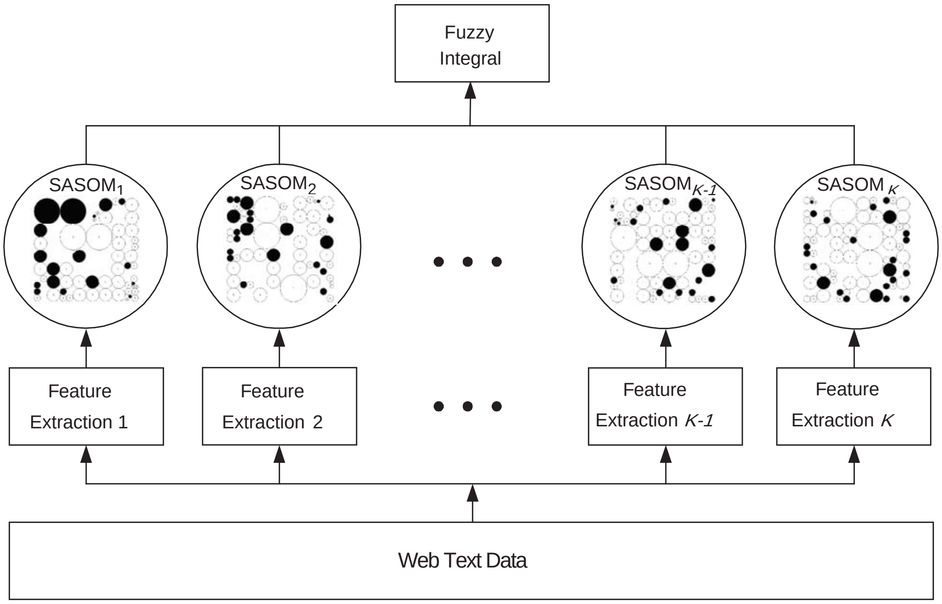


Fig. 1. Overview of the proposed method for web mining.

or “cold.” Problem lies in estimating user’s preference on unknown web page as “hot” or “cold” (“medium” and “cold” are merged because “medium” cases are few). Pazzani reported that naïve Bayes classifier performed better than other methods such as neural network and ID3 for the data [27]. For comparison, majority voting is also adopted. Experimental result indicates that the proposed method using fuzzy integral shows better performance than previous method based on naive Bayes classifier and ensemble of SASOM’s using majority voting.

This paper is organized as follows: Section 2 presents the related works of web mining, fuzzy integral, and SASOM. Section 3 explains the proposed method that uses the ensemble of SASOM’s using fuzzy integral and feature extraction methods. Section 4 shows the experimental results.

2. Related works

Web mining can be categorized into three areas: web content mining, web structure mining, and web usage mining [9]. Web content mining describes the automatic search of information resource available online and involves mining web data content [22]. Web content mining can be seen from two perspectives: information retrieval view and database view [15]. Web structure mining is frequently used to extract the relevance web pages from the web such as Google [10]. Web log analysis is a representative example of web usage mining [38]. There are many applications of web mining and some of them are summarized in Table 1.

Fuzzy integral can be used to combine classifiers and there are applications in vision, face recognition, and handwritten digit recognition. Mirhosseini proposed a novel face recognition system that allowed incorporation of the importance of individual facial components in the recognition task by

Table 1
Summary of web mining research

	Problem	Method	Citation
Web content mining	Feature extraction, Semi-structured DB	Information gain, relational DB	[21,25]
Web structure mining	Cyber community crawling, web page ranking	HITS algorithm, PageRank	[2,17]
Web usage mining	Prediction of customers' purchase behavior	Combination of multiple classifiers	[12]

using fuzzy integral [23]. Cho combined multiple networks based on fuzzy integral and showed the usefulness with the recognition problem of on-line handwriting characters [8]. Pham proposed an algorithm for fusing multiple handwritten-numeral classifiers using fuzzy integral in the sense of adaptive aggregation [29]. Kumar combined neural networks for multispectral data classification using fuzzy integral [16].

One of the big challenges of data mining is the organization and retrieval of documents from archives [24]. Kohonen et al. have demonstrated the utility of a huge self organizing map with more than one million nodes to partition a little less than 7 million patent abstracts where documents are represented by 500-dimensional feature vectors [14]. As an effective pattern recognizer, there are some variants of basic SOM. Cho developed SASOM that can adapt its structure as well as its weights. Bauer et al. presented a growing self-organizing map (GSOM) algorithm. The GSOM has a general hypercubical shape that is adapted during the learning [1]. Suganthan proposed hierarchical overlapped self-organizing maps (HOSOM). HOSOM model has one first level SOM and several partially overlapping second-level SOM's [33].

Syskill & Webert system was designed to help users distinguish interesting web pages on a particular topic from uninteresting ones [27]. They use naïve Bayes classifiers to estimate user profile. Each user starts from initial page and records "like" or "dislike" of the web page by using icon until the number of documents that are checked reaches 10. After 10-recorded pages are collected, Bayes classifier learning algorithm is executed on them. They show that Bayes classifier could perform well with small training set. Syskill & Webert system analyzes web page and recommends links by the preference value of Bayes classifier learned for each user.

Mladenic presented several known and some new feature scoring measures appropriate for feature subset selection on large text data such as information gain, cross entropy, mutual information, weight of evidence, and odds ratio [25]. Lewis discovered the effect of selecting varying numbers and kinds of features for use in predicting category membership on the Reuters and MUC-3 text categorization data sets. In these experiments, optimal feature set size for word-based indexing was found to be low (10–15 features) despite the large training set [20]. Pal proposed soft computing method for feature analysis including various methods using fuzzy logic, neural networks and genetic algorithms for feature ranking, selection and extraction with structure preserving dimensionality reduction [26].

3. Fuzzy integration of SASOM's

In this section, a general framework for estimating user profile using the ensemble of machine learning techniques is proposed. Combination of classifiers can perform better when each classifier is independent to each other. In this purpose, some researchers use different feature sets to train independently and train each classifier using different feature extraction methods [6,33]. Using representative feature extraction methods, each SASOM is trained independently and combined using fuzzy integral. The importance of each classifier is determined subjectively at combination stage.

3.1. Feature selection

Feature selection is the procedure of ranking features based on the information such as frequency and dependency. In text classification, feature is a term in text and has binary value (exist or not). Feature selection procedure is necessary because there are more than 5000 or 6000 features in a collection of 20 web documents. Many features are not useful to improve performance and get difficult to learn classifier.

TFIDF is multiplication of term frequency and inverse document frequency. This measure is frequently used in text retrieval and very simple. For example, suppose that there are 20 documents, among which 10 documents contain "Hello" (DF = 0.5), and 1000 terms in 20 documents where "Hello" occurs 120 times (TF = 0.12). TFIDF of this term is calculated with the following formula (TFIDF = 0.036).

$$\text{TFIDF} = \text{TF} \times \log \frac{1}{\text{DF}}. \tag{1}$$

TFIDF does not use class information of training data to calculate the importance of features. This can degrade the performance of classification. Information gain is the method based on information theory. S is a set of pages and E is expected information gain. $E(W, S)$ means that the expectation of term W on the documents set S.

$$\begin{aligned} E(W,S) &= I(S) - P(W = \text{present}) I(S_{w=\text{present}}) + P(W = \text{absent}) I(S_{w=\text{absent}}), \\ I(S) &= \sum_{c \in \{\text{hot, cold}\}} -p(S_c) \log_2(p(S_c)). \end{aligned} \tag{2}$$

The last feature extraction method is odds ratio [25], which has been used when the goal is to make a good prediction for one of the class values [31].

$$\text{OddsRatio}(W) = \log \frac{\text{odds}(W = \text{present}|C_1)}{\text{odds}(W = \text{present}|C_2)}, \tag{3}$$

where C_1 and C_2 are class labels of binary classification problem. odds(X_i) is defined as follows. X_i is the probability variable such as the probability that term W is in text and class label of the text is C_1.

$$\text{odds}(X_i) = \begin{cases} \dfrac{1/n^2}{1 - 1/n^2}, & P(X_i) = 0, \\ \dfrac{1 - 1/n^2}{1/n^2}, & P(X_i) = 1, \\ \dfrac{P(X_i)}{1 - P(X_i)}, & P(X_i) \neq 0 \wedge P(X_i) \neq 1, \end{cases} \tag{4}$$

where n is a number of examples.

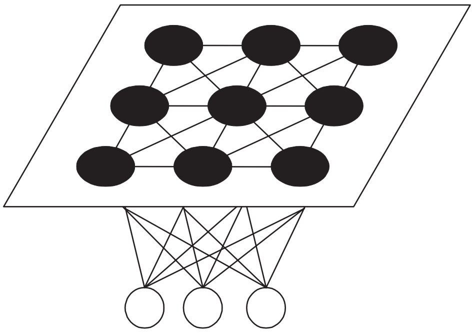


Fig. 2. Basic structure of self-organizing maps.

These three feature selection methods have different properties. TFIDF does not consider class values of documents when calculating the relevance of features while information gain uses class labels of documents. Odds ratio uses class labels of documents but they find useful features to classify only one specific class.

Web documents have tags such as '⟨', '⟩' and non-letters that are not appropriate as feature. They are eliminated from the index. Remaining words can be features for the classification. Each feature has the binary value that means presence or absence of the term. In content mining problem for recommendation, HTML document is classified as "like" or "dislike." For all terms, following statistics are considered:
(1) TF: Term frequency,
(2) DF: Document frequency,
(3) (Term $\in$ Text A),
(4) (Term $\notin$ Text A),
(5) (Text A contains the term) and (Text A is "like"),
(6) (Text A contains the term) and (Text A is "dislike"),
(7) (Text A does not contain the term) and (Text A is "like"),
(8) (Text A does not contain the term) and (Text A is "dislike").

Multiplication of TF and log(1.0/DF) is TFIDF. The set of documents that satisfies (3) is $S_{w=\text{present}}$. Odds ratio uses (5)–(8). Each document is represented with the vector of features that is "0" or "1" and class label. The three different feature extraction methods produce three different representations for each document.

3.2. SASOM

SOM is a neural network model that has property of preserving topology of map and is frequently used to visualize high-dimensional data to low-dimensional space. Fig. 2 shows basic structure of SOM. White node represents input node where input vector is inputted and black node represents neuron. Each neuron competes with other neurons to become the winning node.

Basic SOM fixes the structure of map and shows low performance in classification because each node has data that have different class labels. This property is very useful in unsupervised clustering but becomes weak property in classification. When a node has data that have different class labels, SASOM divides the node into a submap of 4 nodes. Dynamic node splitting scheme is able to simultaneously determine a suitable number of nodes and the connection weights between input and output nodes in self-organizing map.

The basic procedure for SASOM is very simple like this:

(1) Start with a basic SOM (in our case, a 4 × 4 map in which each node is fully connected to all input nodes).
(2) Train the current network with the Kohonen's algorithm [27].
(3) Calibrate the network using known I/O patterns to determine:
    (a) which node should be replaced with a submap of several nodes (in our case, 2 × 2 map), and
    (b) which node should be deleted.
(4) Unless every node represents a unique class, go to step 2.

Basic learning algorithm of SOM is as follows.

$$\|x - w_c\| = \min_i\{\|x - w_i\|\}. \tag{5}$$

The initial map of the network consists of 4 × 4 nodes. The weight vector of node i shall be denoted by $w_i \in R^n$. Euclidean distance between x and w_i is minimized at w_c called winner node. All nodes within N_c update weights by following rule where $\alpha(t)$ is a learning rate, $0 < \alpha(t) < 1$.

$$w_i(t+1) = \begin{cases} w_i(t) + \alpha(t)[x(t) - w_i(t)] & \text{if } i \in N_c(t), \\ w_i(t) & \text{if } i \notin N_c(t). \end{cases} \tag{6}$$

A node representing more than one class is replaced with several nodes. Weights of child nodes of parent node are determined as follows:

$$C = \frac{(P \times 2) + \sum N_c}{S}, \tag{7}$$

where N_c is the neighborhood nodes of child and S is the number of $N_c + 2$.

Fig. 3 shows an instance of node splitting. In this case, the weight of C_0 is determined as follows:

$$C_0 = \frac{(P_4 \times 2) + P_0 + P_1}{4}. \tag{8}$$

Fig. 4 shows how the network structure changes as some nodes representing duplicated classes are replaced by several nodes having finer resolution. The structure of the network is quite similar to Kohonen's self-organizing map except for the irregular connectivity in the map.

Fig. 5 shows the whole procedure of learning independent classifiers using different feature sets. Feature set is a set of terms in a collection of web documents. Using feature selection methods, we can extract the most relevant features, respectively. For example, the first feature subset selects "Initial," "Distance," "Node" and "Several" as the most relevant features. Using these features, input vectors are constructed. For example, the first web documents using the first feature subset has a vector $\langle 0,1,0,1,1\rangle$ that means "Initial" does not exist, "Distance" exists, "Node" does not exist, "Several" exists, and the class label of this document is "1." Using these input vectors, each SASOM is trained.

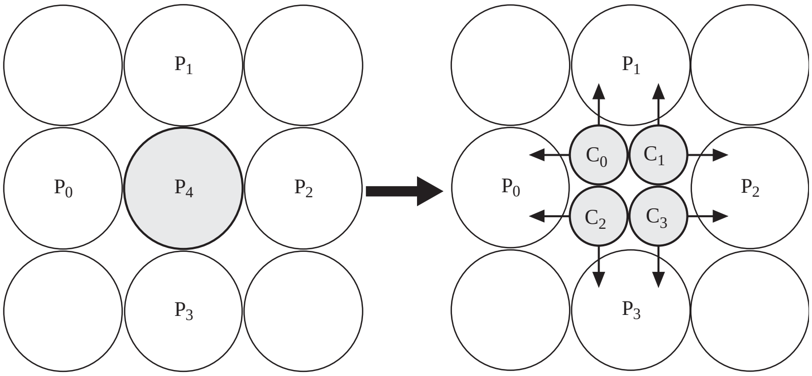


Fig. 3. An instance of node splitting.

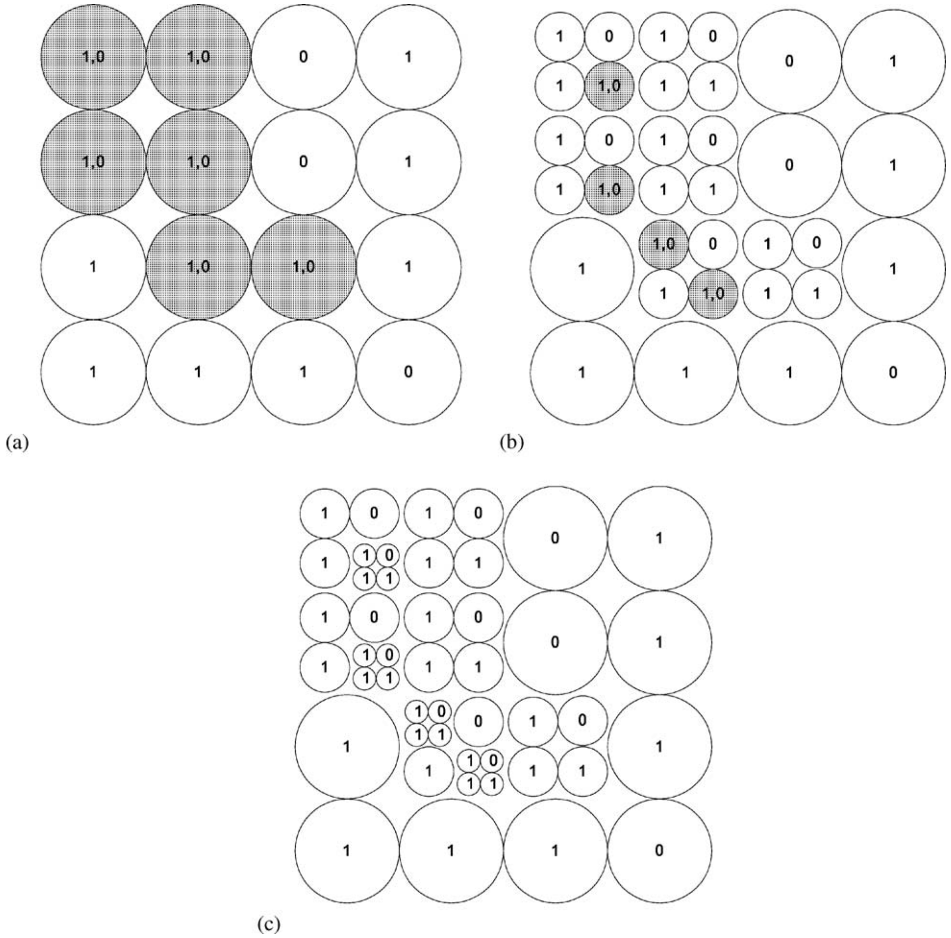


Fig. 4. Map configuration changed through learning. (a) Initial status. (b) Intermediate status. (c) Final status.

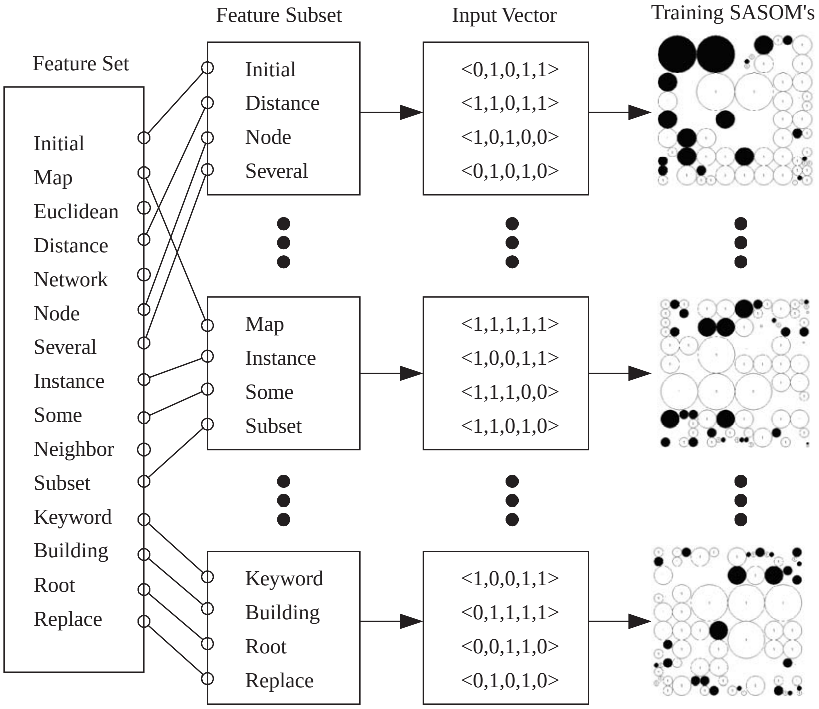


Fig. 5. Training SASOM's using different feature sets.

3.3. Fuzzy integral

There are many combination methods that use the decision of multiple classifiers. They assume that each classifier has the same relevance to the problem. Weighted averaging calculates the relevance of each classifier by using objective measure such as classification performance on training data. Fuzzy integral provides the importance of each classifier measured subjectively. Final decision is integrated from the evidence of classifier for each class and the importance of classifiers subjectively defined by users.

The fuzzy integral introduced by Sugeno [37] and the associated fuzzy measures [19,39] provide a useful way for aggregating information. Fuzzy integral is defined as follows

Definition 1. Let X be a finite set of elements. A set function $g: 2^X \to [0,1]$ with
(1) $g(\varnothing)=0$,
(2) $g(X)=1$,
(3) $g(A) \leqslant g(B)$ if $A \subset B$ is called a fuzzy measure.

Fuzzy measure assigns a real value between 0 and 1 for each subset of X. From the definition of a fuzzy measure g, Sugeno introduced the so-called g_λ-fuzzy measures satisfying the following additional property. For all $A, B \subset X$ and $A \cap B = \varnothing$,

$$g(A \cup B) = g(A) + g(B) + \lambda g(A)g(B) \quad \text{for some } \lambda > -1. \tag{9}$$

It affords that the measure of the union of two disjoint subsets can be directly computed from the component measures.

Sugeno developed the concept of the fuzzy integral, which is a nonlinear functional that is defined with respect to a fuzzy measure, especially g_λ-fuzzy measure.

Definition 2. Let X be a finite set, and $h: X \to [0,1]$ be a fuzzy subset of X. The fuzzy integral over X of the function h with respect to a fuzzy measure g is defined by

$$h(x) \circ g(\cdot) = \max_{E \subseteq X} \left[\min \left(\min_{x \in E} h(x), g(E) \right) \right]. \tag{10}$$

The following properties of the fuzzy integral can be easily proved [10]:
(1) If $h(x) = c$ for all $x \in X, 0 \leqslant c \leqslant 1$, then $h(x) \circ g(\cdot) = c$.
(2) If $h_1(x) \leqslant h_2(x)$ for all $x \in X$, then $h_1(x) \circ g(\cdot) \leqslant h_2(x) \circ g(\cdot)$.
(3) If $\{A_i \mid i = 1, \ldots, n\}$ is a partition of the set X, then

$$h(x) \circ g(\cdot) \geqslant \max_{i=1}^{n} e_i, \tag{11}$$

where e_i is the fuzzy integral of h with respect to g over A_i.

The calculation of the fuzzy integral is as follows: Let $Y = \{y_1, y_2, \ldots, y_n\}$ be a finite set and let $h: Y \to [0,1]$ be a function. Suppose $h(y_1) \geqslant h(y_2) \geqslant h(y_3) \geqslant \cdots \geqslant h(y_n)$. Then a fuzzy integral e with respect to a fuzzy measure g over Y can be computed by

$$e = \max_{i=1}^{n} [\min(h(y_i), g(A_i))], \tag{12}$$

where $A_i = \{y_1, y_2, \ldots, y_i\}$. λ is given by solving the equation

$$\lambda + 1 = \prod_{i=1}^{n} (1 + \lambda g^i), \quad \lambda \in (-1, +\infty) \quad \text{and} \quad \lambda \neq 0. \tag{13}$$

This equation is derived from following recursive calculation. λ can be easily calculated by using the $(n-1)$st degree polynomial.

$$\begin{aligned} &g(A_1) = g(\{y_1\}) = g^1, \\ &g(A_i) = g^i + g(A_{i-1}) + \lambda g^i g(A_{i-1}) \quad \text{for } 1 < i \leqslant n. \end{aligned} \tag{14}$$

Let $C = \{c_1, c_2, c_3, \ldots, c_N\}$ be a set of classes, where binary classification problem has $|C| = 2$. Let $Y = \{y_1, y_2, \ldots, y_n\}$ be a set of classifiers, and $h_k: Y \to [0,1]$ be partial evaluation of the object A (to be classified) for class c_k. $h_k(y_i)$ is an indication of how certain we are in the classification of object A to be in class w_k using the network y_i. The set Y is sorted by the value of $h_k(y_i)$ for each class in descending order. A_{ki} means a set of former i elements in Y for the class k.

$$\text{Final class} = \operatorname*{argmax}_{c_k \in C} \left[\max_{i=1}^{n} [\min(h_k(y_i), g(A_{ki}))] \right]. \tag{15}$$

Each SASOM determines the class label of an unknown document as “0” or “1” (in the binary classification problem). If SASOM$_1$ classifies the document as “0”, $h_0(\text{SASOM}_1) = 1.0$ and

$h_1(\mathrm{SASOM}_1) = 0.0$. Supposed that there are three SASOM's, user evaluates classifiers as g^1, g^2 and g^3, respectively. λ is calculated from g^1, g^2, and g^3. It is easily determined from the 2nd degree polynomial based on (13). For each class k, classifiers are sorted by $h_k(\mathrm{SASOM}_i)$. By the sorted order, they are labeled as y_1, y_2 and y_3. With $g(y_1)$, $g(y_1, y_2)$ and $g(y_1, y_2, y_3)$, the class label of the unknown document is determined using formula (15).

4. Experimental results

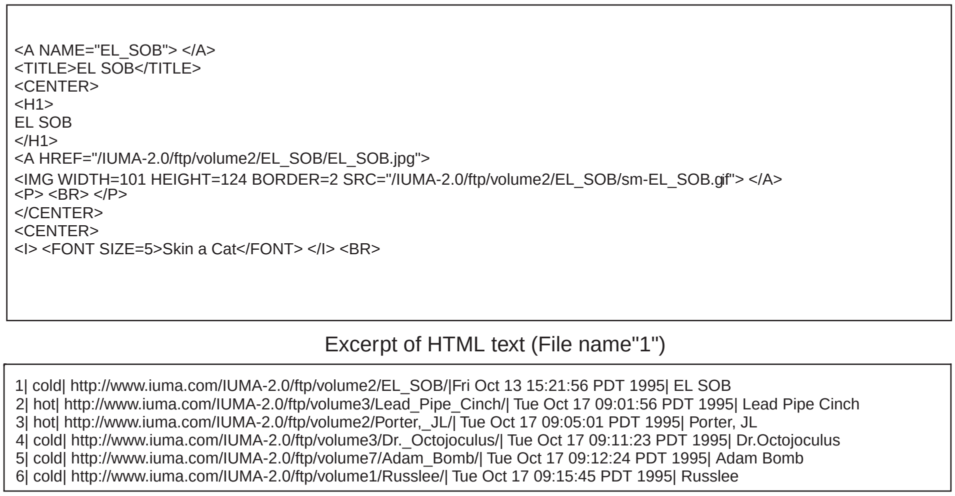
The proposed ensemble aims to estimate user profile by learning the data of web documents. From the UCI KDD database, Syskill & Webert data that have web documents and user's preference value ("hot" or "cold") are accessible. Syskill & Webert data have four different topics "Bands," "Biomedical," "Goats," and "Sheep," among which we use "Goats" and "Bands" data.

"Goats" data have 70 HTML documents and "Bands" 61 HTML documents. Each document has the class label of "hot" or "cold." Fig. 6 shows HTML file and rating data. Each HTML file contains texts related with the topic. Rating file contains file name, rating, URL, date and title orderly. Preprocessing of web documents constructs input vector with selected features and class label. From training data, we extract k important features using three different feature selection methods. Each method ranks all features by different manner. Fig. 7 shows different ranks of features for each method. Using Bands data with 10 training documents, 1200 terms are gathered. In this figure, ranks of a term are different for each method. Document $D = \langle v_1, v_2, v_3, \ldots, v_{128}, c\rangle$ has three different input vectors that are used to train SASOM. The procedure of preprocessing of HTML documents is as follows:

(1) Eliminate non-letters,
(2) Change capital letter to small one; Stop list is a set of following features
    (a) Sort terms by the frequency,
    (b) Select 600 terms that are highly ranked as stop list.
(3) Eliminate stop list,
(4) Make index table ⟨feature, list of documents that have the feature⟩,
(5) Calculate feature relevance using Information gain, TFIDF and odds ratio,
(6) Sort the features by TFIDF and select k features; Sort the features by information gain and select k features; Sort the features by odds ratio and select k features, where k is determined as 128 [27],
(7) Construct input vector for training and test data.

Fig. 7 shows that the three feature extraction methods perform differently. If they are similar to rank terms, the rank of terms are similar but in the figure, the rank of the term for three feature extraction method are very different. In this figure, height of the bar is rank of the term. If the term is relevant, the height of the bar is short because the rank is small.

The problem to solve is to predict unknown documents' classes using known web documents with ensemble of three different SASOMs trained using the input vectors. For each topic, we have conducted 8 different experiments (each experiment has different number of training data and test data). The number of training data is increased by 5 from 10 to 45. Remaining data are used as test set. Experiments are repeated 10 times and the result is the average of them. For comparison, Pazzani's results of naïve Bayes classifier, nearest neighbor, ID3, perceptron, Backpropagation, PEBLS, and Rocchio are used [27].



```
<A NAME="EL_SOB"> </A>
<TITLE>EL SOB</TITLE>
<CENTER>
<H1>
EL SOB
</H1>
<A HREF="/IUMA-2.0/ftp/volume2/EL_SOB/EL_SOB.jpg">
<IMG WIDTH=101 HEIGHT=124 BORDER=2 SRC="/IUMA-2.0/ftp/volume2/EL_SOB/sm-EL_SOB.gif"> </A>
<P> <BR> </P>
</CENTER>
<CENTER>
<I> <FONT SIZE=5>Skin a Cat</FONT> </I> <BR>
```

Excerpt of HTML text (File name"1")

```
1| cold| http://www.iuma.com/IUMA-2.0/ftp/volume2/EL_SOB/|Fri Oct 13 15:21:56 PDT 1995| EL SOB
2| hot| http://www.iuma.com/IUMA-2.0/ftp/volume3/Lead_Pipe_Cinch/| Tue Oct 17 09:01:56 PDT 1995| Lead Pipe Cinch
3| hot| http://www.iuma.com/IUMA-2.0/ftp/volume2/Porter,_JL/| Tue Oct 17 09:05:01 PDT 1995| Porter, JL
4| cold| http://www.iuma.com/IUMA-2.0/ftp/volume3/Dr._Octojoculus/| Tue Oct 17 09:11:23 PDT 1995| Dr.Octojoculus
5| cold| http://www.iuma.com/IUMA-2.0/ftp/volume7/Adam_Bomb/| Tue Oct 17 09:12:24 PDT 1995| Adam Bomb
6| cold| http://www.iuma.com/IUMA-2.0/ftp/volume1/Russlee/| Tue Oct 17 09:15:45 PDT 1995| Russlee
```

Syskill & Webert ratings

Fig. 6. UCI Syskill & Webert data.

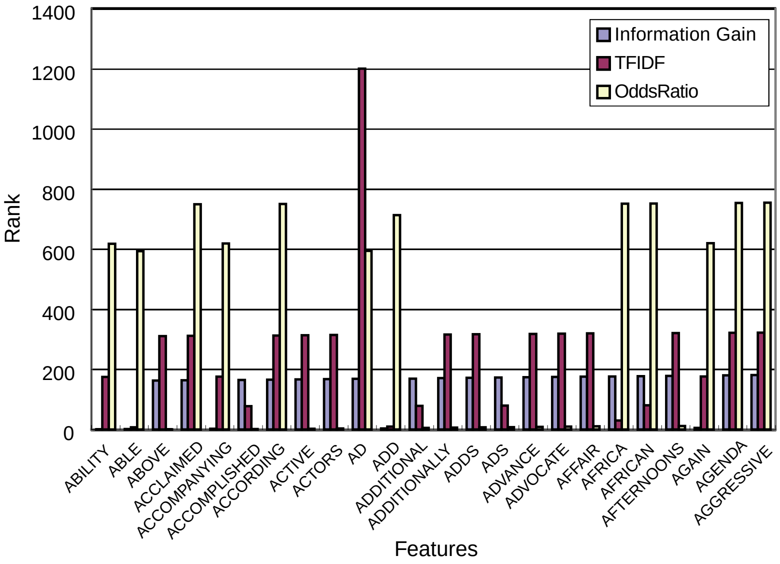


Fig. 7. Rank of features for each feature extraction method (If the height of bar is high, the term is recognized as irrelevant feature in the context of the feature selection method).

Fig. 8 shows the performance of individual classifiers trained using the feature subsets. There are three different feature subsets (information gain, TFIDF and odds ratio). Each classifier shows different performance. In Bands, information gain performs the best. In Goats, odds ratio and TFIDF are good. Accuracy means the ratio of correct predictions. Fig. 9 shows different maps of SASOM's

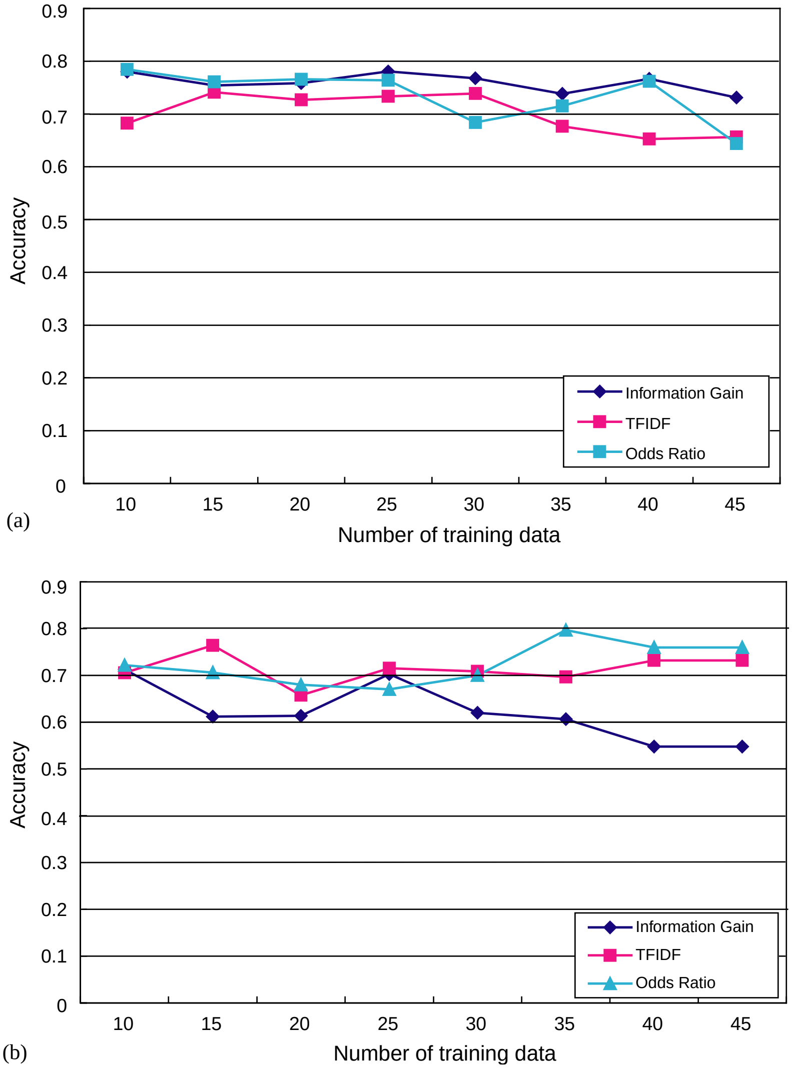


Fig. 8. Performance of single SASOM classifier that is trained using feature subset. (a) Bands. (b) Goats.

for three different feature extraction methods. In Fig. 9(c), there is only nodes with label "1" and this is an exceptional case. Other SASOMs' that use input vectors from odds ratio have nodes with label "1" or "0."

User has several reasons to have selected a web page for "hot." In SOM, neighbor nodes represent similar characteristics though they have different class label. In Fig. 9 (a), there are five distinct groups and we can estimate that user has five different criteria for the decision. Other map analyzes user's criterion with different manner. By analyzing the criterion, we can capture conceptual understanding of user's classification and apply it to web mining. Table 2 summarizes the number of nodes in the topology of SASOM.

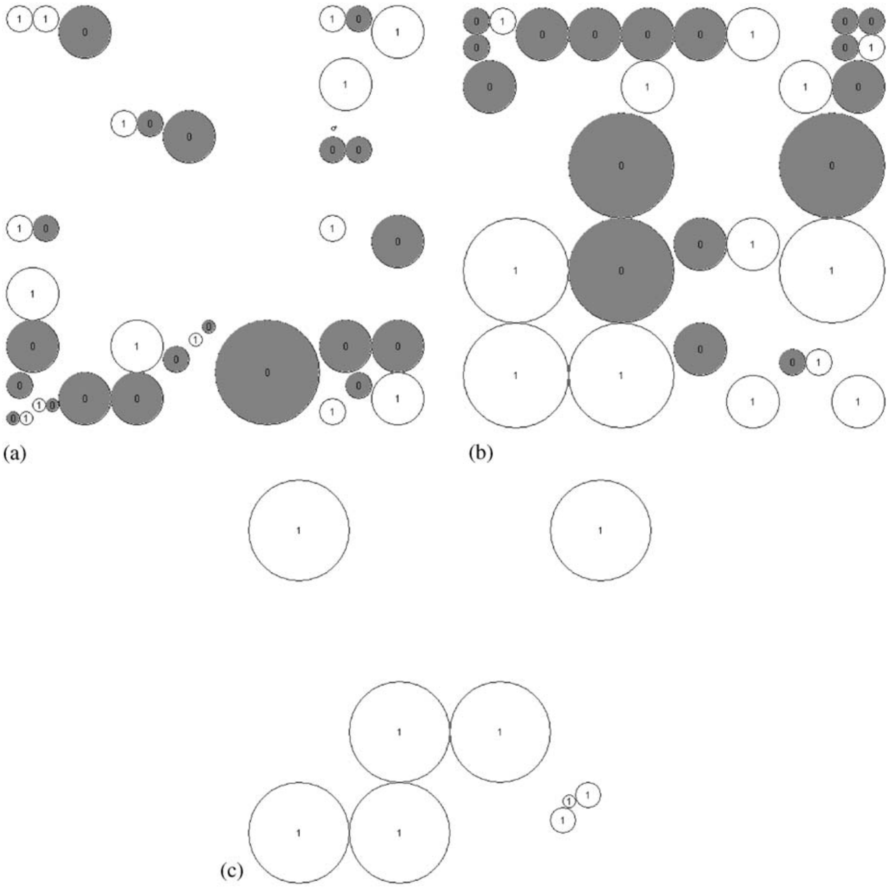


Fig. 9. Topology of maps in Goats data (The number of training data is 45). (a) Information gain. (b) TFIDF. (c) Odds ratio.

Table 2
The number of nodes in the topology of SASOM (Bands data)

The number of training data	Information gain	TFIDF	Odds ratio
10	9	8	8
15	12	11	4
20	11	11	4
25	17	20	5
30	25	21	5
35	19	25	5
40	24	26	5
45	30	27	5

Table 3
Fuzzy measure of classifiers (1: information gain, 2: TFIDF, 3: Odds ratio)

	Bands			Goats		
The number of training data	g^1	g^2	g^3	g^1	g^2	g^3
10	0.99	0.99	0.99	0.99	0.01	0.99
15	0.99	0.99	0.99	0.99	0.99	0.01
20	0.99	0.99	0.99	0.99	0.99	0.01
25	0.99	0.99	0.99	0.99	0.01	0.99
30	0.99	0.99	0.01	0.99	0.99	0.99
35	0.99	0.01	0.99	0.99	0.99	0.99
40	0.99	0.99	0.99	0.99	0.99	0.99
45	0.99	0.99	0.00	0.99	0.01	0.99

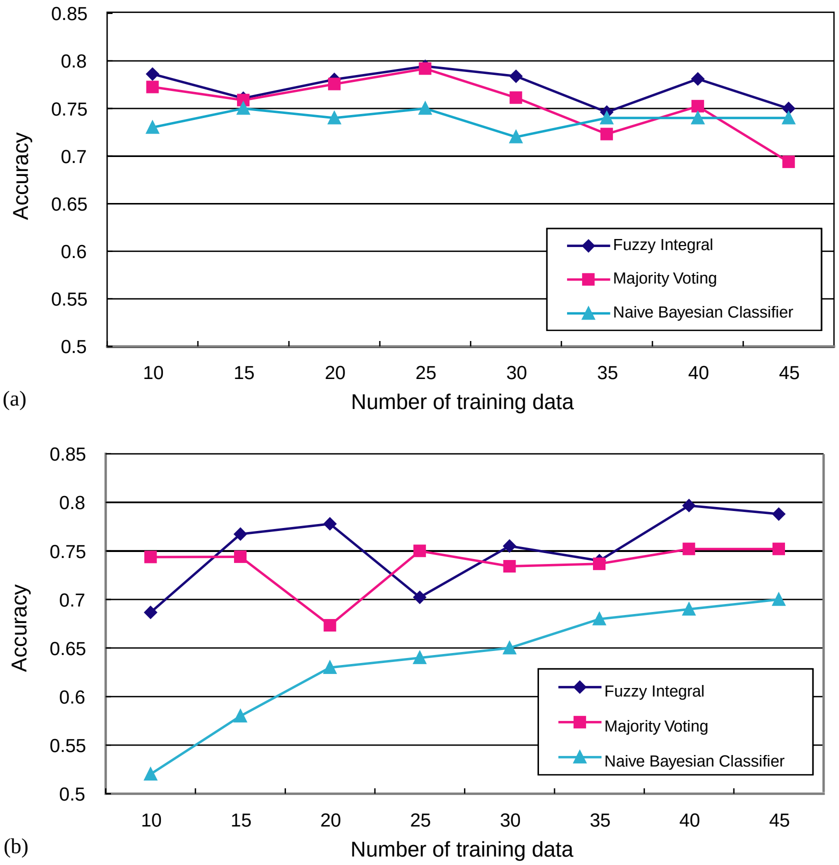


Fig. 10. Combination of SASOM's using fuzzy integral. (a) Bands. (b) Goats.

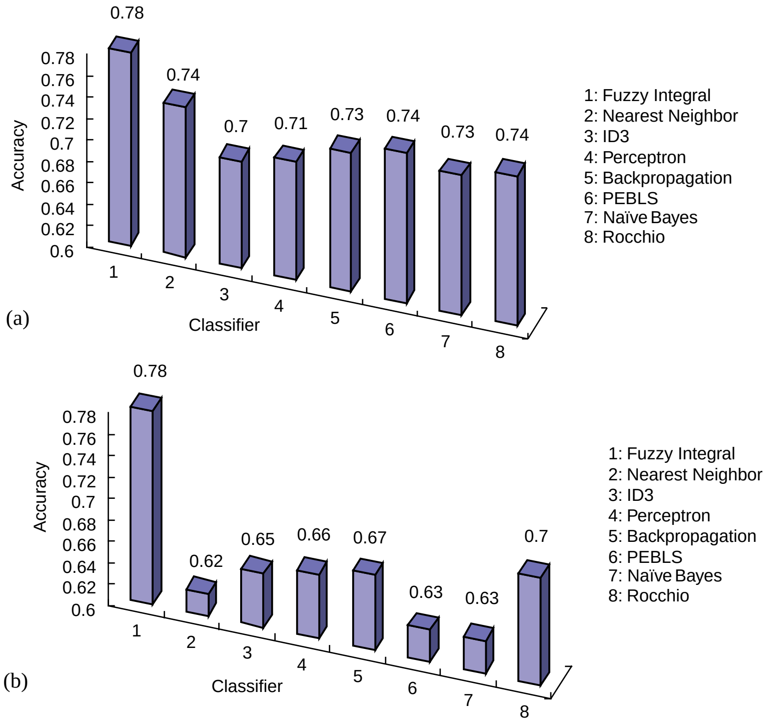


Fig. 11. Performance comparison with other classifiers. (a) Bands. (b) Goats.

Table 3 shows fuzzy measure of classifiers that are determined subjectively. Subjective assignment is a strong point of the proposed method, and some strategies based on the preference on the validation data set have been proposed previously [7], even though we could not use it this time due to the lack of the validation data set. Also, we were not that careful to adjust the measures and just assigned the rough values based on the performance on the training data set.

Fig. 10 shows the performance of ensemble of SASOM's using fuzzy integral and majority voting. For comparison, results of Pazzani's naïve Bayes classifier are used. In Bands, fuzzy integral performs better than naive Bayes classifier and majority voting. In Goats, fuzzy integral performs better than other two methods. Majority voting of SASOM's shows better performance than naive Bayes classifier but lower performance than fuzzy integral.

Implementation of the proposed method needs efficient indexing method because there are many terms in the text collections. Feature ranking needs additional information such as term occurrence, document frequency of the term, and conditional probability information. Empirically, SASOM does not need much time to learn the user profile because the size of training data is not large in this problem.

Fig. 11 shows a comparison with other seven classifiers. Twenty examples are chosen as a reasonable intermediate number of examples. Training set size is 20 and the remaining of data is used as a test set. It can be confirmed that fuzzy integral performs better than other classifiers.

5. Concluding remarks

In this paper, we have proposed an ensemble of SASOM's using fuzzy integral to classify web documents based on user's preference. Experimental results show that the proposed method performs better than the results of previous studies and majority voting of SASOM's. Fuzzy integral provides the method of measuring the importance of classifiers subjectively. SASOM can classify documents with high performance and visualize its map to understand internal mechanism. The proposed method can be effectively applied to web content mining for predicting user's preference as user profile.

Acknowledgements

This work was supported by Biometrics Engineering Research Center, and Brain Science and Engineering Research Program sponsored by Korean Ministry of Science and Technology.

References

[1] H.-U. Bauer, T. Villmann, Growing a hypercubical output space in a self-organising feature map, IEEE Trans. Neural Networks 8 (2) (1997) 218–226.

[2] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Comput. Networks ISDN Syst. 30 (1–7) (1998) 107–117.

[3] S. Chakrabarti, Mining the Web: Analysis of Hypertext and Semi Structured Data, Morgan Kaufmann Publishers, Los Angeles, 2002.

[4] S.-B. Cho, Neural-network classifiers for recognizing totally unconstrained handwritten numerals, IEEE Trans. Neural Networks 8 (1) (1997) 43–53.

[5] S.-B. Cho, Self-organizing map with dynamical node splitting: application to handwritten digit recognition, Neural Comput. 9 (6) (1997) 1343–1353.

[6] S.-B. Cho, Ensemble of structure-adaptive self-organizing maps for high performance classification, Inform. Sci. 123 (1–2) (2000) 103–114.

[7] S.-B. Cho, J.-H. Kim, Combining multiple neural networks by fuzzy integral for robust classification, IEEE Trans. Syst. Man Cybern. 25 (2) (1995) 380–384.

[8] S.-B. Cho, J.-H. Kim, Multiple network fusion using fuzzy logic, IEEE Trans. Neural Networks 6 (2) (1995) 497–501.

[9] R. Cooley, B. Mobasher, J. Srivastava, Web mining: information and pattern discovery on the world wide web, Proc. 9th IEEE Internat. Conf. on Tools with Artificial Intelligence (ICTAI'97), 1997, pp. 558–567.

[10] Google, http://www.google.com.

[11] S. Hettich, S.D. Bay, The UCI KDD Archive, http://kdd.ics.uci.edu.

[12] E. Kim, W. Kim, Y. Lee, Combination of multiple classifiers for the customer's purchase behavior prediction, Dec. Support Syst. 34 (2) (2003) 167–175.

[13] T. Kohonen, The self-organizing map, Proc. IEEE 78 (9) (1990) 1464–1480.

[14] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, A. Saarela, Self organization of a massive document collection, IEEE Trans. Neural Networks 11 (2000) 574–585.

[15] R. Kosala, H. Blockeel, Web mining research: a survey, SIGKDD Explorations 2 (1) (2000) 1–15.

[16] A.S. Kumar, S.K. Basu, K.L. Majumdar, Robust classification of multispectral data using multiple neural networks and fuzzy integral, IEEE Trans. Geosci. Remote Sensing 35 (3) (1997) 787–790.

[17] R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, Trawling the web for emerging cyber-communities, Comput. Networks 31 (11–16) (1999) 1481–1493.

[18] O.-W. Kwon, J.-H. Lee, Text categorization based *k*-nearest neighbor approach for web site classification, Inform. Process. Manage. 39 (1) (2003) 25–44.

[19] K. Leszeynski, P. Penczek, W. Grochulskki, Sugeno's fuzzy measures and fuzzy clustering, Fuzzy Sets and Systems 15 (1985) 147–158.

[20] D. Lewis, Feature selection and feature extraction for text categorization, Proc. DARPA Workshop on Speech and Natural Language, 1992, pp. 212–217.

[21] W.-S. Li, J. Shim, K.S. Candan, WebDB: a system for querying semi-structured data on the web, J. Visual Language Comput. 13 (1) (2002) 3–33.

[22] S.K. Madria, S.S. Rhowmich, W.K. Ng, F.P. Lim, Research issues in web data mining, Proc. 1st Internat. Conf. on Data Warehousing and Knowledge Discovery, 1999, pp. 303–312.

[23] A.R. Mirhosseini, H. Yan, K.-M. Lam, T. Pham, Human face image recognition: an evidence aggregation approach, Comput. Vis. Image Understanding 71 (2) (1998) 213–230.

[24] S. Mitra, S.K. Pal, P. Mitra, Data mining in soft computing framework: a survey, IEEE Trans. Neural Networks 13 (1) (2002) 3–14.

[25] D. Mladenic, M. Grobelnik, Feature selection on hierarchy of web documents, Dec. Support Syst. 35 (2003) 45–87.

[26] N.R. Pal, Soft computing for feature analysis, Fuzzy Sets and Systems 103 (1999) 201–221.

[27] M. Pazzani, D. Billsus, Learning and revising user profiles: the identification of interesting web sites, Mach. Learning 27 (1997) 313–331.

[28] M. Pazzani, J. Muramatsu, D. Billsus, Syskill & Webert: identifying interesting web sites, Proc. 13th Nat. Conf. on AI (AAAI 96), Portland, OR, 1996, pp. 54–61.

[29] T.D. Pham, Combination of multiple classifiers using adaptive fuzzy integral, Proc. 2002 IEEE Internat. Conf. on Artificial Intelligence Systems (ICAIS'02), 2002, pp. 50–55.

[30] J.R. Quinlan, Induction of decision trees, Mach. Learning 1 (1986) 81–106.

[31] C.J.V. Rijsbergen, D.J. Harper, M.F. Porter, The selection of good search terms, Inform. Process. Manage. 17 (1981) 77–91.

[32] J. Srivastava, R. Cooley, M. Deshpande, P. Tan, Web usage mining: discovery and application of web usage patterns from web data, SIGKDD Explorations 1 (2) (2000) 12–23.

[33] P.N. Sugathan, Hierarchical overlapped SOM-based multiple classifiers combination, Proc. 5th Internat. Conf. on Control, Automation, Robotics & Vision (ICARCV'98), 1998, pp. 924–927.

[34] P.N. Suganthan, Pattern classification using multiple hierarchical overlapped self-organising maps, Pattern Recog. 34 (11) (2001) 2173–2179.

[35] M. Sugeno, Fuzzy measures and fuzzy integrals : a survey. Fuzzy Automata and Decision Processes, North Holland, Amsterdam, 1977, pp. 89–102.

[36] A. Verikas, A. Lipnickas, K. Malmqvist, M. Bacauskiene, A. Gelzinis, Soft combination of neural classifiers: a comparative study, Pattern Recog. Lett. 20 (1999) 429–444.

[37] J. Vesanto, SOM-based data visualization methods, Intell. Data Anal. 3 (2) (1999) 111–126.

[38] Y. Xiao, M.H. Dunham, Efficient mining of traversal patterns, Data Knowledge Eng. 39 (2) (2001) 191–214.

[39] R.R. Yager, Element selection from a fuzzy subset using the fuzzy integral, IEEE Trans. Syst. Man Cybern. 23 (1993) 467–477.