DATA ANALYSIS

DATA CLASSIFICATION

What is data analysis? Nolan (1) gives a definition that is a way of making sense of the patterns that are in, or can be imposed on, sets of figures. In concrete terms, data analysis consists of an observation and an investigation of the given data, and the derivation of characteristics from the data. Such characteristics, or features as they are sometimes called, contribute to the insight of the nature of data. Mathematically, the features can be regarded as some variables, and the data are modeled as a realization of these variables with some appropriate sets of values. In traditional data analysis (2), the values of the variables are usually numerical and may be transformed into symbolic representation. There are two general types of variables: discrete and continuous. Discrete variables vary in units, such as the number of words in a document or the population in a region. In contrast, continuous variables can vary in less than a unit to a certain degree of accuracy. The stock price and the height of people are examples of this type. The suitable method for collecting values of discrete variables is counting, and for continuous ones it is measurement.

The task of data analysis is required among various application fields, such as agriculture, biology, economics, government, industry, medicine, military, psychology, and science. The source data provided for different purposes may be in various forms, such as text, image, or wave form. There are several basic types of purposes for data analysis:

1. Obtain the implicit structure of data
2. Derive the classification of data
3. Search particular objects in data

For example, the stockbroker would like to get the future trend of the stock price, the biologist needs to divide animals into taxonomies, and the physician tries to find the related symptoms of a given disease. The techniques to accomplish these purposes are generally drawn from statistics that provide well-defined mathematical models and probability laws. In addition, some theories, such as fuzzy-set theory, are also useful for data analysis in particular. This article is an attempt to give a brief description of these techniques and concepts of data analysis. In the following section, a variety of data analysis methods are introduced and illustrated by examples. We first give two categories of data analysis according to its initial conditions and resultant uses. Next, we show two well known methods based on different mathematical models. In the second section, an approach to data analysis for Internet applications is proposed. Some improvements of the data analysis methods are discussed in the third section. Finally, we give a brief summary of this article.

DATA ANALYSIS METHODS

In data analysis, the goals are to find significant patterns in the data and apply this knowledge to some applications. Analysis is generally performed in the following stages:

1. Feature selection
2. Data classification
3. Conclusion evaluation

The first stage consists of the selection of the features in the data according to some criteria. For instance, features of people may include their height, skin color, and fingerprints. Considering the effectiveness of human recognition, the fingerprint, which is the least ambiguous, may get the highest priority for selection. In the second stage, the data are classified according to the selected features. If the data consist of at least two features, e.g., the height and the weight of people, which can be plotted in a suitable coordinate system, we can inspect so-called scatter plots and detect clusters or contours for data grouping. Furthermore, we can investigate ways to express data similarity. In the final stage, the conclusions drawn from the data would be compared with the actual demands. A set of mathematical models has been developed for this evaluation. In the following sections, we first divide the study of data analysis into two categories according to different initial conditions and resultant uses. Then, we introduce two famous models for data analysis. Each method will be discussed first, followed by examples. Because the feature selection depends on the actual representations of data, we postpone the discussion about this stage until the next section. In this section, we focus on the classification procedure based on the given features.

A Categorization of Data Analysis

There are a variety of ways to categorize the methods of data analysis. According to the initial conditions andthe resultant uses, there are two categories, supervised data analysis and unsupervised data analysis. The term *supervised* means that human knowledge has to be provided for the process. In supervised data analysis, we specify a set of classes called a *classification template* and select some samples from the data for each class. These samples are then labeled by the names of the associated classes. Based on this initial condition, we can automatically classify the other data termed *to-be-classified* data. In *unsupervised* data analysis, there is no classification template, and the resultant classes depend on the samples. Following are descriptions of supervised and unsupervised data analysis with an emphasis on their differences.

Supervised Data Analysis. The classification template and the well-chosen samples are given as an initial state and contribute to the high accuracy of data classification. Consider the K nearest-neighbor classifier, which is a typical example of supervised data analysis. The input to the classifier includes a set of labeled samples S, a constant value K, and a to-be-classified datum X. The output after the classification is a label denoting a class to which X belongs. The classification procedure is as follows.

1. Find the K nearest neighbors (K NNs) of X from S.
2. Choose the dominant classes by K NNs.

J. Webster (ed.), *Wiley Encyclopedia of Electrical and Electronics Engineering*. Copyright © 2007 John Wiley & Sons, Inc.

3. If there exists only one dominant class, label X by this class; otherwise, label X by any dominant class.
4. Add X to S, and the process terminates.

The first step selects K samples from S such that the values of the selected features (also called patterns) of these K samples are closest to those of X. Such a similarity may be expressed in a variety of ways. The measurement of distances among the patterns is one of the suitable instruments, for example, the Euclidean distance as shown in Eq. (1). Suppose the K samples belong to a set of classes; the second step is to find the set of dominant classes C'. A dominant class is a class that contains the majority of the K samples. If there is only one element in C', say class C_i, we assign X to C_i. On the other hand, if C' contains more than one element, X is assigned to an arbitrary class in C'. After deciding on the class of X, we label it and add it into the set S.

$$\delta(X,Y) = \sqrt{\sum_{k=1}^{m}(X_k - Y_k)^2}, \qquad (1)$$

where each datum is represented by m features.

Example. Suppose there is a dataset about the salaries and ages of people. Table 1 gives such a set of samples S and the corresponding labels. There are three labels that denote three classes: rich, fair, and poor. These classes are determined based on the assumption that richness depends on the values of the salary and age. In Table 1, we also append the rules for assigning labels for each age. From the above, we can get the set membership of each class.

$$C_{\text{rich}} = \{Y_1,Y_4,Y_8\}, \qquad C_{\text{fair}} = \{Y_2,Y_5,Y_6,Y_{10}\},$$
$$C_{\text{poor}} = \{Y_3,Y_7,Y_9\}$$

If there is a to-be-classified datum X with age 26 and salary \$35,000 ($35k$), we apply the classification procedure to classify it. Here we let the value of K be 4 and use the Euclidean distance as the similarity measure.

1. The set of 4 NNs is $\{Y_4,Y_5,Y_6,Y_9\}$.
2. The dominant class is the class C_{fair} because $Y_6, Y_5 \in C_{\text{fair}}$, $Y_4 \in C_{\text{rich}}$, and $Y_9 \in C_{\text{poor}}$.
3. Label X by C_{fair}.
4. New sample S contains an updated class $C_{\text{fair}} = \{Y_2,Y_5,Y_6,Y_{10}, X\}$.

We can also give an assumed rule to decide the corresponding label for the age of X as shown in Table 1. Obviously, the conclusion drawn from the above classification coincides with such an assumption from human knowledge.

Unsupervised Data Analysis. Under some circumstances, data analysis consists of a partition of the whole data set into a number of subsets. Moreover, the data within each subset have to be similar to a high degree, whereas the data between different subsets have to be similar to a very low degree. Such subsets are called clusters, and the way to find a good partition is sometimes also called cluster analysis. There are a variety of methods developed to handle this problem. A common characteristic among them is the iterative nature of the algorithms.

The C-mean clustering algorithm is representative in this field. The input contains the sample set S and a given value C, which denotes the number of clusters in the final partition. Notice that no labels are assigned to the samples in S in advance. Before classification, we must specify an initial partition W_0 with C clusters. The algorithm terminates when it converges to a stable situation in which the current partition remains the same as the previous one. Different initial partitions can lead to different final results. One way to get the best partition is to apply this algorithm with all different W_0's. To simplify the illustration, we only consider a given W_0 and a fixed C. The classification procedure is as follows.

1. Let W be W_0 on S.
2. Compute the mean of each cluster in W.
3. Evaluate the nearest mean of each sample and move a sample if its current cluster is not the one corresponding to its nearest mean.
4. If any movement occurs, go to step 2; otherwise, the process terminates.

The first step sets the current partition W to be W_0. Then we compute a set of means M in W. In general, a mean is a virtual sample representing the whole cluster. It is straightforward to use averaging as the way to find M. Next, we measure the similarities between each sample in S and every mean M. Suppose a sample Y_j belongs to a cluster C_i in the previous partition W, while another cluster C_k has a mean nearest to Y_j. Then we move Y_j from C_i to C_k. Finally, if there exists such a sample movement, the partition W would become a new one and requires more iterations. On the other hand, if no such movement occurs during an iteration, the partition would become stable and produce the final clustering.

Example. Consider the data in Table 1 again. Suppose there is no label on each sample and only the salary and the age data are used as the features for analysis. For clarity, we use a pair of values on the two features to represent a sample, for instance, the pair (20, 25k) refers to the sample Y_1. Suppose there is an initial partition containing two clusters C_1 and C_2. Let the means of these clusters be M_1 and M_2, respectively. The following shows the iterations for the clustering.

1. For the initial partition W: $C_1 = \{Y_1,Y_2,Y_3,Y_4,Y_5\}$, $C_2 = \{Y_6,Y_7,Y_8,Y_9,Y_{10}\}$.

The first iteration

1. <label>2.</label>M_1 = (23.6, 24k), M_2 = (33.6, 44k).
2. <label>3.</label>Move Y_4 from C_1 to C_2; move Y_7 and Y_9 from C_2 to C_1.
3. <label>4.</label>For the new partition W: $C_1 = \{Y_1,Y_2,Y_3,Y_5,Y_7,Y_9\}$, $C_2 = \{Y_4,Y_6,Y_8,Y_{10}\}$.

The second iteration <list1 type="custom">

Table 1. A Set of Samples with the Salary and Age Data

Sample	Age	Salary	Label	Assumed rules to assign labels
Y_1	20	25k	Rich	rich, >20k; poor, <10k
Y_2	22	15k	Fair	rich, >26k; poor, <13k
Y_3	24	15k	Poor	rich, >35k; poor, <16k
Y_4	24	40k	Rich	
Y_5	28	25k	Fair	rich, >44k; poor, <22k
Y_6	30	40k	Fair	rich, >50k; poor, <25k
Y_7	30	20k	Poor	
Y_8	32	60k	Rich	rich, >56k; poor, <28k
Y_9	36	30k	Poor	rich, >68k; poor, <34k
Y_{10}	40	70k	Fair	rich, >80k; poor, <40k
X	26	35k	Fair	rich, >38k; poor, <19k

1. <label>2.</label>M_1 = (26.6, 21.6k), M_2 = (31.5, 52.5k).
2. <label>3.4.</label>There is no sample movement; the process terminates.

We can easily find a simple discriminant rule behind this final partition. All the samples with salaries lower than 40k belong to C_1, and the others belong to C_2. Hence we may conclude with a discriminant rule that divides S into two clusters by checking the salary data. If we use another initial partition, say W', where $C_1 = \{Y_1,Y_3,Y_5,Y_7,Y_9\}$ and $C_2 = \{Y_2,Y_4,Y_6,Y_8,Y_{10}\}$, the conclusion is the same. The following process yields another partition with three clusters.

1. For the initial partition W: $C_1 = \{Y_1,Y_4,Y_7\}$, $C_2 = \{Y_2,Y_5,Y_8\}$, $C_3 = \{Y_3,Y_6,Y_9,Y_{10}\}$.

The first iteration

1. <label>2.</label>M_1 = (24.6, 28.3k), M_2 = (27.3, 33.3k), M_3 = (32.5, 38.7k)
2. <label>3.</label>Move Y_4 from C_1 to C_2, move Y_2 and Y_5 from C_2 to C_1, move Y_8 from C_2 to C_3, move Y_3 from C_3 to C_1, move Y_9 from C_3 to C_2.
3. <label>4.</label>For the new partition W: $C_1 = \{Y_1,Y_2,Y_3,Y_5,Y_7\}$, $C_2 = \{Y_4,Y_9\}$, $C_3 = \{Y_6,Y_8,Y_{10}\}$

The second iteration

1. <label>2.</label>M_1 = (24.8, 20k), M_2 = (30, 35k), M_3 = (34, 56.6k).
2. <label>3.</label>Move Y_6 from C_3 to C_2.
3. <label>4.</label>For the new partition W: $C_1 = \{Y_1,Y_2,Y_3,Y_5,Y_7\}$, $C_2 = \{Y_4,Y_6,Y_9\}$, $C_3 = \{Y_8,Y_{10}\}$.

The third iteration

1. <label>2.</label>M_1 = (24.8, 20k), M_2 = (30, 36.6k), M_3 = (36, 65k).
2. <label>3.4.</label>There is no sample movement; the process terminates.

After three iterations, we have a stable partition and also conclude with the discriminant rule that all the samples with salaries lower than 30k belong to C_1, the other samples with salaries lower than 60k belong to C_2, and the remainder belongs to C_3. The total number of iterations depends on the initial partition, the number of clusters, the given features, and the similarity measure.

Methods for Data Analysis

In the following, we introduce two famous techniques for data analysis. One is Bayesian data analysis based on probability theory, and the other is fuzzy data analysis based on fuzzy-set theory.

Bayesian Data Analysis. Bayesian inference, as defined In Ref. 3, is the process of fitting a probability model to a set of samples, which results in a probability distribution to make predictions for to-be-classified data. In this environment, a set of samples is given in advance and labeled by their associated classes. Observing the patterns contained in these samples, we can obtain not only the distributions of samples for the classes but also the distributions of samples for the patterns. Therefore, we can compute a distribution of classes for these patterns and use this distribution to predict the classes for the to-be-classified data based on their patterns. A typical process of Bayesian data analysis contains the following stages:

1. Compute the distributions from the set of labeled samples.
2. Derive the distribution of classes for the patterns.
3. Evaluate the effectiveness of these distributions.

Suppose a sample containing the pattern a on some features is labeled class C_i. First, we compute a set of probabilities $P(C_i)$ that denote a distribution of samples for different classes and let each $P(a|C_i)$ denote the conditional probability of a sample containing the pattern a, given that the sample belongs to the class C_i. In the second stage, the conditional probability of a sample belonging to class C_i, given that the sample contains the pattern a, can be formulated as follows:

$$P(C_i|a) = \frac{P(a|C_i)P(C_i)}{P(a)}, \tag{2}$$

where

$$P(a) = \sum_i P(a|C_i)P(C_i)$$

From Eq. (3), we can derive the probabilities of a sample belonging to classes according to the patterns contained in the sample. Finally, we can find a way to determine the class by using these probabilities. The following is a simple illustration of data analysis based on this probabilistic technique.

Example. Consider the data in Table 1. We first gather the statistics and transform the continuous values into discrete ones as in Table 2. Here we have two discrete levels, young and old, representing the age data, and three levels, low, median, and high, referring to the salary data. We collect all the probabilities and derive the ones for prediction based on Eq. (3).

$$\begin{aligned}
&P(\text{young, low}|C_{\text{rich}}) = \tfrac{1}{3}, \qquad P(\text{young, low}|C_{\text{fair}}) = \tfrac{1}{2},\\
&P(\text{young, low}|C_{\text{poor}}) = \tfrac{1}{3},\\
&P(\text{young, median}|C_{\text{rich}}) = \tfrac{1}{3}, \qquad P(\text{young, median}|C_{\text{fair}}) = 0,\\
&P(\text{young, median}|C_{\text{poor}}) = 0, \ldots\\
&P(\text{young, low}) = \tfrac{4}{10}, \qquad P(\text{young, median}) = \tfrac{1}{10},\\
&P(\text{young, high}) = 0, \ldots\\
&P(C_{\text{rich}}) = \tfrac{3}{10}, \qquad P(C_{\text{fair}}) = \tfrac{2}{5}, \qquad P(C_{\text{poor}}) = \tfrac{3}{10}\\
&P(C_{\text{rich}}|\text{young, low}) = \tfrac{1}{4}, \qquad P(C_{\text{fair}}|\text{young, low}) = \tfrac{1}{2},\\
&P(C_{\text{poor}}|\text{young, low}) = \tfrac{1}{4},\\
&P(C_{\text{rich}}|\text{young, median}) = 1, \qquad P(C_{\text{fair}}|\text{young, median}) = 0,\\
&P(C_{\text{poor}}|\text{young, median}) = 0, \ldots
\end{aligned}$$

Because there are two features representing the data, we compute the joint probabilities instead of the individual probabilities. Here we assume that the two features have the same degree of significance. At this point, we have constructed a model to express the data with their two features. The derived probabilities can be regarded as a set of rules to decide the class of any to-be-classied datum.

If there is a to-be-classified datum X whose age is 26 and salary is 35k, we apply the derived rules to label X. We transform the pattern of X to indicate that the age is young and the salary is low. To find the suitable rules, we can define a penalty function $\lambda(C_i|C_j)$, which denotes the payment when a datum belonging to C_j is classified into C_i. Let the value of this function be 1 if C_j is not equal to C_i and 0 if two classes are the same. Furthermore, we can define a distance measure $\iota(X, C_i)$ as in Eq. (5), which represents the total amount of payments when we classify X into C_i. We conclude that the lower the value of $\iota(X, C_i)$, the higher the probability that X belongs to C_i. In this example, we label X by C_{fair} because $\iota(X, C_{\text{fair}})$ is the lowest.

$$\iota(X, C_i) = \sum_j \lambda(C_i|C_j)P(C_j|X) \tag{3}$$

$$\iota(X, C_{\text{rich}}) = 0 \times \tfrac{1}{4} + 1 \times \tfrac{1}{2} + 1 \times \tfrac{1}{4} = \tfrac{3}{4}$$

$$\iota(X, C_{\text{fair}}) = \tfrac{1}{2}, \qquad \iota(X, C_{\text{poor}}) = \tfrac{3}{4}$$

Fuzzy Data Analysis. Fuzzy set theory, established by Zadeh (4), allows a gradual membership $MF_A(X)$ for any datum X on a specified set A. Such an approach more adequately models the data uncertainty than using the common notion of set membership. Take cluster analysis as an example. Each datum belongs to exactly one cluster after the classification procedure. Often, however, the data cannot be assigned exactly to one cluster in the real world, such as the jobs of a busy person, the interests of a researcher, or the conditions of the weather. In the following, we replace the previous example for supervised data analysis with the fuzzy-set notion to show its characteristic.

Consider a universe of data U and a subset A of U. Set theory allows to express the membership of A on U by the characteristic function $F_A(X)$:$U \rightarrow \{0,1\}$.

$$F_A(X) = \begin{cases} 1, & X \in A \\ 0, & X \notin A \end{cases} \tag{4}$$

From the above, it can be clearly determined whether X is an element of A or not. However, many real-world phenomena make such a unique decision impossible. In this case, expressing in of membership is more suitable. A fuzzy set A on U can be represented by the set of pairs that describe the membership function $MF_A(X)$:$U \rightarrow [0,1]$ as defined In Ref. 5.

$$A = \{(X, MF_A(X)) | X \in U,\ MF_A(X) \in [0,1]\} \tag{5}$$

Example. Table 3 contains a fuzzy-set representation of the dataset in Table 1. The membership function of each sample is expressed in a form of possibility that stands for the degree of the acceptance that a sample belongs to a class. Under the case of supervised data analysis, the to-be-classified datum X needs to be labeled using an appropriate classification procedure. All the distances between each sample and X are calculated using the two features and Euclidean distance.

1. Find the K nearest neighbors (K NNs) of X from S.
2. Compute the membership function of X for each class.
3. Label X by the class with a maximal membership.
4. Add X to S and stop the process.

The first stage in finding K samples with minimal distances is the same, so we have the same set of four nearest neighbors $\{Y_4, Y_5, Y_6, Y_9\}$ when the value of $K = 4$. Let $\delta(X, Y_j)$ denote the distance between X and the sample Y_j. In the next stage, we calculate the membership function $MF_{C_i}(X)$

Table 2. A Summary of Probability Distribution for the Data in Table 1

Sample	Rich	Fair	Poor	Expressions of new condensed features
Young	2	2	1	Age is lower than 30
Old	1	2	2	Other ages
Low	1	2	3	Salary is lower than 36k
Median	1	1	0	Other salaries
High	1	1	0	Salary is higher than 50k

Table 3. Fuzzy-set Membership Functions for the Data in Table 1

Sample	Rich	Fair	Poor	Estimated distances between the sample and X
Y_1	0.5	0.2	0.3	11.66
Y_2	0.1	0.5	0.4	20.39
Y_3	0	0.2	0.8	20.09
Y_4	0.6	0.3	0.1	5.38
Y_5	0.2	0.5	0.3	10.19
Y_6	0.2	0.5	0.2	6.4
Y_7	0	0	1	15.52
Y_8	0.9	0.1	0	25.7
Y_9	0	0.3	0.7	11.18
Y_{10}	0.4	0.6	0	37.69
X	0.2	0.42	0.38	

of X for each class C_i as follows:

$$MF_{C_i}(X) = \frac{\sum_j MF_{C_i}(Y_j)\delta(X,Y_j)}{\sum_j \delta(X,Y_j)} \quad \forall Y_j \in k \text{ NNs of } X \tag{6}$$

$$MF_{C_{\text{rich}}}(X) = \frac{0.6 \times 5.38 + 0.2 \times 10.19 + 0.2 \times 6.4 + 0 \times 11.18}{5.38 + 10.19 + 6.4 + 11.18} \approx 0.2$$

$$MF_{C_{\text{fair}}}(X) = \frac{0.3 \times 5.38 + 0.5 \times 10.19 + 0.6 \times 6.4 + 0.3 \times 11.18}{5.38 + 10.19 + 6.4 + 11.18} \approx 0.42$$

$$MF_{C_{\text{poor}}}(X) = \frac{0.1 \times 5.38 + 0.3 \times 10.19 + 0.2 \times 6.4 + 0.7 \times 11.18}{5.38 + 10.19 + 6.4 + 11.18} \approx 0.38$$

Because the membership of X for class C_{fair} is higher than all others, we label X by C_{fair}. The resultant membership directly gives a confidence measure of the classification.

DATA ANALYSIS ON INTERNET DATA

The dramatic growth of information systems over the past years has brought about the rapid accumulation of data and an increasing need for information sharing. The World Wide Web (*WWW*) combines the technologies of the uniform resource locator (*URL*) and hypertext to organize the resources in the Internet into a distributed hypertext system (5). As more and more users and servers register on the WWW, data analysis on its rich content is expected to produce useful results for various applications. Many research communities such as network management (5), information retrieval (5), and database management (5) have been working in this field.

Many tools for Internet resource discovery (6) use the results of data analysis on the WWW to help users find the correct positions of the desired resources. However, many of these tools essentially keep a keyword-based index of the available resources (Web pages). Owing to the imprecise relationship between the semantics of keywords and the Web pages (7), this approach clearly does not fit the user requests well. From the experiments in (7), the text-based classifier that is 87

The goal of Internet data analysis is to derive a classification of a large amount of data, which can provide a valuable guide for the WWW users. Here the data are the Web pages produced by the information providers of the WWW. In some cases, data about the browsing behaviors of the WWW users are also interesting to the data analyzers, such as the most popular sites browsed or the relations among the sites in a sequence of browsing. Johnson and Fotouhi (8) propose a technique to aid users to roam through the hypertext environment. They gather and analyze all the browsing paths of some users to generate a summary as a guide for other users. Many efforts have been made to apply the results of such data analysis (8). In this article, we focus on the Web pages that are the core data of the WWW. First, we present a study on the nature of Internet data. Then we show the feature selection stage and enforce a classification procedure to group the data at the end.

Each site within the Web environment contains one or more Web pages. Under this environment, any WWW user can make a request to any site for any Web page in it. Moreover, the user can also roam through the Web by means of the anchor information provided in each Web page. Such an approach has resulted in several essential difficulties for data analysis.

1. Huge amounts of data
2. Frequent changes
3. Heterogeneous presentations

Basically the Internet data originate from all over the world, and the amount of data is huge. As any WWW user can create, delete, and update the data, and change the locations of the data at any time, it is difficult to get a precise view of the data. Furthermore, the various forms of expressing the same data also reveal the status of the chaos on the WWW. As a whole, Internet data analysis should be able to handle the large amount of data and control the uncertainty factors in a practical way. The data analysis procedure consists of the following stages:

1. Observe the data.
2. Collect the samples.
3. Select the features.
4. Classify the data.
5. Evaluate the results.

In the first stage, we observe the data and conclude with a set of features that may be effective for classifying the data. Next, we collect a set of samples based on a given scope. In the third stage, we estimate the fitness of each feature for the collected samples to determine a set of effective features. Then, we classify the to-be-classified data according to the similarity measure on the selected features. At last, we evaluate the classified results and find a way for further improvement.

Data Observation

In the following, we provide two directions for observing the data.

Semantic Analysis. We may consider the semantics of a Web page as potential features. Keywords contained in a Web page can be analyzed to determine the semantics such as which fields it belongs to or what concepts it provides. There have been many efforts at developing techniques to derive the semantics of a Web page. The research results of information retrieval (9, 10) can also be applied for this purpose.

Observing the data formats of Web pages, we can find several parts expressing the semantics of the Web pages to some extent. For example, the title of a Web page usually refers to a general concept of the Web page. An anchor, which is constructed by the home-page designer, provides a URL of another Web page and makes a connection between the two Web pages. As far as the home-page designer is concerned, the anchor texts must sufficiently express the semantics of the whole Web page to which the anchor points. As to the viewpoint of a WWW user, the motivation to follow an anchor is based on the fact that this anchor expresses desired semantics for the user. Therefore, we can make a proper connection between the user's interests and those truly relevant Web pages. We can group the anchor texts to generate a corresponding classification of the Web pages pointed to by these anchor texts. Through this classification we can relieve the WWW users of the difficulties on Internet resource discovery through a query facility.

Syntactic Analysis. Syntactic analysis is based on the syntax of the Web pages to derive a rough classification. Because the data formats of Web pages follow the standards provided on the WWW, for example, hypertext markup language (*HTML*), we can find potential features among the Web pages. Consider the features shown in Table 4. The white pages, which mean the Web pages with a list of URLs, can be distinguished from the ordinary Web pages by a large number of anchors and the short distances between two adjacent anchors within a Web pages. Note that here the distance between two anchors means the number of characters between them. For publication, the set of the headings has to contain some specified keywords, such as "bibliography" or "Publications." The average distance between two adjacent anchors has to be lower than a given threshold and the placement of anchors has to center to the bottom of the Web page.

According to these features, some conclusions may be drawn in the form of classification rules. For instance, the Web page is designed for publication if it satisfies the requirements of the corresponding features. Obviously, this approach is effective only when the degree of support for such rules is high enough. Selection of effective features is a way to improve the precision of syntactic analysis.

Sample Collection

It is impossible to collect all the Web pages, and thus choosing a set of representative samples becomes a very important task. On the Internet, we have two approaches to gather these samples.

1. Supervised sampling
2. Unsupervised sampling

Supervised sampling means the sampling process is based on human knowledge which specifies the scope of the samples. In supervised data analysis, there exists a classification template that consists of a set of classes. The sampling scope can be set based on the template. The sampling is more effective when all classes of the template contain at least one sample. On the other hand, we consider unsupervised sampling if there is not enough knowledge about the scope, as in the case of unsupervised data analysis. The most trivial way to get samples is to choose any subset of Web pages. However, this arbitrary sampling may not fit the requirement of random sampling well. We recommend the use of search engines that provide different kinds of Web pages in a form of directory.

Table 4. Potential Features for Some Kinds of Web Pages

Kind of home page	Potential feature
White page	Number of anchors, average distance between two adjacent anchors
Publication	Headings, average distance between two adjacent anchors, anchor position
Person	Title, URL directory
Resource	Title, URL filename

Feature Selection

In addition to collecting enough samples, we have to select suitable features for the subsequent classification. No matter how good the classification scheme is, the accuracy of the results would not be satisfactory without effective features. A measure for the effectiveness of a feature is to estimate the degree of class separability. A better feature implies a higher class separability. This measure can be formulated as a criterion to select effective features.

Example. Consider the samples shown in Table 5. From Table 4, there are two potential features for white pages, the number of anchors (F_0) and the average distance between two adjacent anchors (F_1). We assume that $F_0 \geq 30$ and $F_1 \leq 3$ when the sample is a white page. However, a sample may actually belong to the class of white pages although it does not satisfy the assumed conditions. For example, Y_6 is a white page although its $F_0 < 30$. Therefore, we need to find a way to select effective features.

From the labels, the set membership of the two classes is as follows, where the class C_1 refers to the class of white pages.

$$C_0 = \{Y_1, Y_2, Y_3, Y_4, Y_5\}, \qquad C_1 = \{Y_6, Y_7, Y_8, Y_9, Y_{10}\}$$

We can begin to formulate the class separability. In the following formula, we assume that the number of classes is c, the number of samples within class C_{j} is n_{j}, and $Y^{\mathrm{i}}{}_{\mathrm{k}}$ denotes the kth sample in class C_{i}. First, we define the interclass separability D_{b}, which represents the ability of a feature to distinguish the data between two classes. Next, we define the intraclass separability D_{w}, which expresses the power of a feature to separate the data within the same class. The two measures are formulated in Eqs. (10) and (8) based on the Euclidean distance defined in Eq. (1)anwar. Since a feature with larger D_{b} and smaller D_{w} implies a better class separability, we define a simple criterion function D_{F_j} [Eq. (12)] as a composition of D_{b} and D_{w} to evaluate the effectiveness of a feature F_{j}. Based on this criterion function, we get $D_{F_0} = 1.98$ and $D_{F_1} = 8.78$. Therefore, F_1 is more effective than F_0 due to its higher class separability.

$$D_{\mathrm{b}} = \frac{1}{2}\sum_{i=1}^{c} P_i \sum_{j\neq i} P_j \frac{1}{n_i n_j} \sum_{k=1}^{n_i}\sum_{m=1}^{n_j} \delta(Y_k^i, Y_m^j) \tag{7}$$

where

$$P_i = \frac{n_i}{\sum_{j=1}^{c} n_j}$$

$$D_{\mathrm{w}} = \frac{1}{2}\sum_{i=1}^{c} P_i \sum_{j=i} P_j \frac{1}{n_i n_j} \sum_{k=1}^{n_i}\sum_{m=1}^{n_j} \delta(Y_k^i, Y_m^j) \tag{8}$$

where

$$P_i = \frac{n_i}{\sum_{j=1}^{c} n_j}$$

$$D_{F_j} = D_{\mathrm{b}} - D_{\mathrm{w}} \tag{9}$$

We have several ways to choose the most effective set of features:

1. Ranking approach
2. Top-down approach
3. Bottom-up approach
4. Mixture approach

Ranking approach selects the features one by one according to the rank of their effectiveness. Each time we include a new feature from the rank, we compute the joint effectiveness of the features selected so far by Eqs. (10)–(12). When the effectiveness degenerates, the process terminates. Using a top-down approach, we consider all the features as the initial selection and drop the features one by one until the effectiveness degenerates. On the contrary, the bottom-up approach adds a feature at each iteration. The worse case of the above two approaches occurs if we choose the bad features earlier in the bottom-up approach or the good features earlier in the top-down approach. The last approach allows us to add and drop the features at each iteration by combining the above two approaches. After determining the set of effective features, we can start the classification process.

Data Classification

In the following, we only consider the anchor semantics as the feature, which is based on the dependency between an anchor and the Web page to which the anchor points. As mentioned previously, the semantics expressed by the anchor implies the semantics of the Web page to which the anchor points, and also describes the desired Web pages for the users. Therefore, grouping the semantics of the anchors is equivalent to classifying the Web pages into different classes. The classification procedure consists of the following stages:

Table 5. A Set of Samples with Two Features. The Labels Come from Human Knowledge

Sample	F_0^a	F_1^b	White page
Y_1	8	5	no
Y_2	15	3.5	no
Y_3	25	2.5	no
Y_4	35	4	no
Y_5	50	10	no
Y_6	20	2	yes
Y_7	25	1	yes
Y_8	40	2	yes
Y_9	50	2	yes
Y_{10}	80	8	yes

[a] F_0 denotes the number of anchors.

[b] F_1 denotes the average distance for two adjacent anchors.

1. Label all sample pages.
2. For each labeled pages, group the texts of the anchors pointing to it.
3. Record the texts of the anchors pointing to the to-be-classified page.
4. Classify the to-be-classified page based on the anchor information.
5. Refine the classification process.

In the beginning, we label all the samples and record all the anchors pointing to them. Then we group together the anchor texts contained in the anchors pointing to the same sample. In the third stage, we group the anchor texts contained in the anchors pointing to the to-be-classified page. After the grouping, we determine the class of the to-be-classified page according to the corresponding anchor texts. At last, we can further improve the effectiveness of the classification process. There are two important measures during the classification process. One is the similarity measure of two data, and the other is the criterion for relevance feedback.

Similarity Measure. After the grouping of samples, we have to measure the degree of membership between the to-be-classified page and each class. Considering the Euclidean distance again, there are three kinds of approaches for such measurement:

1. Nearest-neighbor approach
2. Farthest-neighbor approach
3. Mean approach

The first approach finds the the sample in each class nearest to the to-be-classified page. Among these representative samples, we can choose the class containing the one with a minimal distance and assign the page to it. On the other hand, we can also find the farthest sample in each class from the page. Then we assign the page to the class that contains the representative sample with a minimal distance. The last approach is to take the mean of each class into consideration. As in the previous approaches, the mean of each class represents a whole class, and the one with a minimal distance from the page would be chosen. An example follows by using the mean approach.

Example. Inspect the data shown in Table 6. There are several Web pages and anchor texts contained in some anchors pointing to the Web pages. Here we consider six types of anchor texts, T_1,T_2, . . . ,T_6. The value of an anchor text for a Web page stands for the number of the anchors pointing to the Web page, which contain the anchor text. The labeling is the same as in the previous example. We can calculate the means of the two classes:

$$\boldsymbol{M}_0 = (0,4,1,1,1,0.2,1), \qquad \boldsymbol{M}_1 = (4.2,3.4,2.6,1.4,2,1.4)$$

Suppose there is a Web page X to be classified as shown in Table 6. We can compute the distances between X and the two means. They are $\delta(X, M_0)$ = 6.94 and $\delta(X, M_1)$ = 4.72. Thus we assign X to class C_1.

Relevance Feedback. The set of samples may be enlarged after a successful classification by including the classified Web pages. However, the distance between a to-be-classified page and the nearest mean may be very large, which means that the current classification process does not work well on this Web page. In this case, we reject the classification of such a Web page and wait until more anchor texts for this Web page are accumulated. This kind of rejection not only expresses the extent of the current ability to classify Web pages, but also promotes the precision of the classified results. Furthermore, by the concept of class separability formulated in Eqs. (10)–(12), we can define a similar criterion function D_{S} to evaluate the performance of the current set of samples.

$$D_{\mathbf{S}} = D_{\mathbf{F}}(S) \tag{10}$$

where F is the set of all effective features and S is the current set of samples.

Example. Reconsider the data shown in Table 6. Before we assign X to C_1, the initial D_{S} = 0.75. When C_1 contains X, $D_{\mathrm{S}\cup\{X\}}$ yields a smaller value 0.16. On the other hand, $D_{\mathrm{S}\cup\{X\}}$ becomes 1.26 if we assign X to C_0. Hence, although X is labeled C_1, it is not suitable to become a new sample for the subsequent classification. The set of samples can be

Table 6. A Set of Web Pages with Corresponding Anchor Texts and Labels. The Labels Come from Human Knowledge

Sample	T_1[a]	T_2[b]	T_3[c]	T_4[d]	T_5[e]	T_6[f]	White page
Y_1	0	0	0	1	1	2	no
Y_2	0	1	2	0	0	2	no
Y_3	0	2	0	4	0	0	no
Y_4	0	0	3	0	0	1	no
Y_5	2	2	0	0	0	0	no
Y_6	1	3	0	0	2	3	yes
Y_7	3	3	1	6	3	0	yes
Y_8	4	2	5	0	1	0	yes
Y_9	5	5	3	0	0	2	yes
Y_{10}	8	4	4	1	4	2	yes
X	5	2	0	0	5	0	yes

[a] T_1 = "list."
[b] T_2 = "directory."
[c] T_3 = "classification."
[d] T_4 = "bookmark."
[e] T_5 = "hot."
[f] T_6 = "resource."

enlarged only when such an addition of new samples gains a larger D_S value, which means the class separability is improved.

IMPROVEMENT OF THE DATA ANALYSIS METHODS

Although the previous procedures are able to fit the requirements of data analysis well, there are still problems, such as speed or memory requirements and the complex nature of real-world data. We have to use some heuristic techniques to improve the classification performance. For example, the number of clusters given in unsupervised data analysis has significant impact on the time spent at each iteration and the quality of the final partition. Notice that the initial partition may contribute to a specific sequence of adjustments and then a particular solution. Therefore, we have to find an ideal number of clusters during the analysis according to the given initial partition. The bottom-up approach for decreasing the number of clusters at each iteration is a way to determine the final partition. Given a threshold of similarity among the clusters, we can merge two clusters that are similar enough to become a new single cluster at each iteration. We can find a suitable number of clusters when there are no more similar clusters to be merged. In the following sections, we introduce two more techniques to improve the work of data analysis.

Rough-Set Based Data Analysis

The approach to classifying Internet data by anchor semantics requires a large amount of anchor texts. These anchor texts may be contained in the anchors pointing to the Web pages in different classes. An anchor text is said to be indiscernible when it cannot be used to distinguish the Web pages in different classes. We employ the rough-set theory (11, 12) to find the indiscernible anchor texts, which will then be removed. The remaining anchor texts will contribute to a higher degree of accuracy for the subsequent classification. In addition, the cost of distance computation can also be reduced. In the following, we introduce the basic idea of the rough-set theory and an example for the reduction of anchor texts.

Rough-set Theory. By the rough-set theory, an information system is modeled in the form of a 4-tuple (U,A,V,F), where U represents a finite set of objects, A refers to a finite set of attributes, V is the union of all the domains of the attributes in A, and F is a binary function $(U \times A{:} \rightarrow V)$. The attribute set A often consists of two subsets, one refers to condition attributes $\bar{C}$ and the other stands for decision attributes $\bar{D}$. In the approach of classification on Internet data, U stands for all the Web pages, A is the union of the anchor texts $(\bar{C})$ and the class of Web pages $(\bar{D})$ V is the union of all the domains of the attributes in A, and F handles the mappings. Let B be a subset of A. A binary relation called indiscernibility relation is defined as

$$\mathbf{IND}_B = \{(X_i, X_j) \in U \times U \mid \forall p \in B,\ p(X_i) = p(X_j)\} \qquad \mathbf{(11)}$$

That is, X_i and X_j are indiscernible by the set of attributes B if $p(X_i)$ is equal to $p(X_j)$ for every attribute p in B. IND_B is an equivalence relation that produces an equivalence class denoted $[X_i]_B$ for each sample X_i. With regard to the Internet data, two Web pages X_i and X_j, which have the same statistics for each anchor text in $\bar{C}$ belong to the same equivalence class $[X_i]\bar{C}$ (or $[X_j]\bar{C}$). Let U' be a subset of U. A lower approximation $\mathrm{LOW}_{B,U'}$, which contains all the samples in each equivalence class $[X_i]_B$ contained in U', is defined as

$$\mathbf{LOW}_{B,U'} = \{X_i \in U \mid [X_i]_B \subset U'\} \qquad \mathbf{(12)}$$

Based on Eq. (16), LOW $\bar{C},[X_i]\bar{D}$ contains the Web pages in the equivalence classes produced by IND $\bar{C}$, and these equivalence classes are contained in $[X_i]\bar{D}$ for a given X_i. A positive region POS $\bar{C},\bar{D}$ is defined as the union of LOW $\bar{C},[X_i]$ $\bar{D}$ for each equivalence class produced by IND $\bar{D}$. POS $\bar{D},\bar{D}$ refers to the samples that belong to the same

Table 7. A Set of Data in Symbolic Values Transformed from Table 6

Sample	T_1	T_2	T_3	T_4	T_5	T_6	White page
Y_1	L[a]	L	L	L	L	L	no
Y_2	L	L	L	L	L	L	no
Y_3	L	L	L	M[b]	L	L	no
Y_4	L	L	M	L	L	L	no
Y_5	L	L	L	L	L	L	no
Y_6	L	M	L	L	L	M	yes
Y_7	M	M	L	H[c]	M	L	yes
Y_8	M	L	M	L	L	L	yes
Y_9	M	M	M	L	L	L	yes
Y_{10}	H	M	M	L	M	L	yes
X	M	L	L	L	M	L	yes

[a] L = [0, 2].
[b] M = [3, 5].
[c] H = [6, 8].

class when they have the same anchor texts. As defined In Ref. 13, $\bar{C}$ is independent of $\bar{D}$if each subset $\bar{C}_\mathrm{i}$ in $\bar{C}$ satisfies the criterion that POS $\bar{C}$, $\neq$ POS $\bar{C}_\mathrm{i}$,$\bar{D}$; otherwise, $\bar{C}$ is said to be dependent on $\bar{D}$ The degree of dependency γ $\bar{C}$,$\bar{D}$ is defined as

$$\gamma_{\bar{C},\bar{D}} = \frac{\mathrm{card}(\mathrm{POS}_{\bar{C},\bar{D}})}{\mathrm{card}(U)} \qquad (13)$$

where card denotes set cardinality;

$$\mathrm{CON}_{p,\gamma_{\bar{C},\bar{D}}} = \gamma_{\bar{C},\bar{D}} - \gamma_{\bar{C}-\{p\},\bar{D}} \qquad (14)$$

From these equations, we define the contribution CON$_{\mathrm{p},\gamma}$ $\bar{C}$,$\bar{D}$ of an anchor text p in $\bar{C}$ to the degree of dependency γ $\bar{C}$,CIDbar; by using Eq. (18). According to Eq. (17), we say an anchor text p is dispensable if γ $\bar{C}-\{p\}$,$\bar{D}$ = γ $\bar{C}$,$\bar{D}$. That is, the anchor text p makes no contribution to γ $\bar{C}$,$\bar{D}$ and the value of CON$_{\mathrm{p},\gamma}$ $\bar{C}$,$\bar{D}$ equals 0. The set of indispensable anchor texts is the core of the reduced set of anchor texts. The remaining task is to find a minimal subset of $\bar{C}$ called a reduct of $\bar{C}$ which satisfies Eq. (19) and the condition that the minimal subset is independent of $\bar{D}$.

$$\mathrm{POS}_{\bar{C},\bar{D}} = \mathrm{POS}_{\mathrm{minimalsubsetof}\bar{C},\bar{D}} \qquad (15)$$

Reduction of Anchor Texts. To employ the concepts of the rough-set theory for the reduction of anchor texts, we transform the data shown in Table 6 into those in Table 7. The numerical value of each anchor text is transformed into a symbol according to the range in which the value falls. For instance, a value in the range between 0 and 2 is transformed into the symbol L. This process is a generalization technique usually used for a large database.

By Eq. (18), we can compute CON$_{\mathrm{p},\gamma}$ $\bar{C}$,$\bar{D}$ for each anchor text p and sort them in ascending order. In this case, all CON$_{\mathrm{p},\gamma}$ $\bar{C}$,$\bar{D}$ are 0 except CON$_{\mathrm{T1},\gamma}$ $\bar{C}$,$\bar{D}$. That is, only the anchor text T_1 is indispensable, which becomes the unique core of $\bar{C}$ Next, we use a heuristic method to find a reduct of $\bar{C}$ because such a task has been proved to be NP-complete In Ref. 14. Based on an arbitrary ordering of the dispensable anchor texts, we check the first anchor text to see whether it is dispensable. If it is, then remove it and continue to check the second anchor text. This process continues until no more anchor texts can be removed.

Example. Suppose we sort the dispensable anchor texts as the sequence $\{T_2,T_3,T_4,T_5,T_6\}$, we then check one at a time to see whether it is dispensable. At last, we obtain the reduct $\{T_1,T_6\}$. During the classification process, we only consider these two anchor texts for similarity measure. Let the symbols used in each anchor text be transformed into three discrete values, 0, 1, and 2. The means of the two classes are M_0 = (0, 0) and M_1 = (1, 0.8). Therefore, we classify X into the class C_1 due to its minimum distance. When we use the reduct $\{T_1,T_6\}$ to classify data, the class separability $D_{\{T1},T_{6\}}$ is 0.22. Different reducts may result in different values of class separability. For instance, the class separability becomes 0.27 if we choose the reduct $\{T_1,T_2\}$.

Hierarchical Data Analysis

Consider the 1-nearest-neighbor classifier for supervised data analysis. We may not want to compute all the distances each time a to-be-classified datum X arrives. We can organize the set of samples into a hierarchy of subsets and record a mean M_i for each subset S_i and the farthest distance d_i from M_i to any sample in S_i. If there exists a nearest neighbor of X in a subset other than S_i, we do not need to compute the distances between X and all the samples in S_i as the triangular inequality [Eq. (20)] holds. Such techniques can reduce the computation time to find the nearest neighbor.

$$\delta(X, M_i) - d_i \geq \delta(X, Y) \qquad (16)$$

where Y is the nearest neighbor of X.

SUMMARY

In this article, we describe the techniques and concepts of data analysis. A variety of data analysis methods are introduced and illustrated by examples. Two categories, supervised data analysis and unsupervised data analysis, are presented according to their different initial conditions and resultant uses. Two methods for data analysis are also described, which are based on probability theory and fuzzy-set theory, respectively. An approach of data analysis Internet data is presented. Improvements for the data analysis methods are also discussed.

BIBLIOGRAPHY

1. B. Nolan *Data Analysis: An Introduction*, Cambridge, UK: Polity Press, 1994.
2. J. W. Tukey *Exploratory Data Analysis*, Reading, MA: Addison-Wesley, 1977.
3. A. Gelman *et al. Bayesian Data Analysis*, London: Chapman & Hall, 1995.
4. L. A. Zadeh Fuzzy sets, *Information Control*, **8**: 338–353, 1965.
5. H. Bandemer W. Nather *Fuzzy Data Analysis*, Dordrecht: Kluwer, 1992. T. Berners-Lee, R. Cailliau, A. Luotonen, H. F. Nielsen, and A. Secret, The world wide web, *Communications of the ACM*, **37**(8): 76–82, 1994. M. Baentsch, L. Baum, G. Molter, S. Rothkugel, and P. Sturm, Enhancing the web's infrastructure: from caching to replication, *IEEE Internet Computing*, **1**(2): 18–27,March/April 1997. V. N. Gudivada, V. V. Raghavan, W. I. Grosky, and R. Kasanagottu, Information retrieval on the world wide web, *IEEE Internet Computing*, **1**(5): 58–68,September/October 1997. D. Florescu, A. Levy, and A. Mendelzon, Database techniques for the world wide web: A Survey, *ACM SIGMOD Record*, **27**(3): 59–74,September 1998.
6. K. Obraczka P. B. Danzig S. H. Li Internet resource discovery services, *IEEE Comput. Mag.*, **26** (9): 8–22, 1993.
7. C. S. Chang A. L. P. Chen Supporting conceptual and neighborhood queries on WWW, *IEEE Trans. Syst. Man Cybernet.* in press. S. Chakrabarti, B. Dom, and P. Indyk, Enhanced hypertext categorization using hyperlinks, Proceedings of ACM SIGMOD Conference on Management of Data, pp. 307–318, 1998.
8. A. Johnson F. Fotouhi Automatic touring in hypertext systems, Proc. IEEE Phoenix Conf. Comput. Commun., Phoenix, 1993, 524–530. A. Buchner and M. D. Mulvenna, Discovering internet marketing intelligence through online analytical web usage mining, *ACM SIGMOD Record*, **27**(4): 54–61,December 1998. T. W. Yan, M. Jacobsen, H. Garcia-Molina, and U. Dayal, From User Access patterns to dynamic hypertext linking, *Computer Networks and ISDN Systems*, **28**: 1007–1014, 1996.
9. G. Salton M. J. McGill *Introduction to Modern Information Retrieval*, New York: McGraw-Hill, 1983.
10. G. Salton *Automatic Text Processing*, Reading, MA: Addison Wesley, 1989.
11. Z. Pawlak Rough Set, *Commun. ACM*, **38** (11): 88–95, 1995.
12. Z. Pawlak *Rough Sets: Theoretical Aspects of Reasoning about Knowledge*, Norwell, MA: Kluwer, 1991.
13. X. Hu N. Cercone Mining knowledge rules from databases: A rough set approach, Proc. 12th Int. Conf. Data Eng., Ed. Stanley Y. W. Su, New Orleans, LA, 1996, 96–105.
14. R. Slowinski (Editor) *Handbook of Applications and Advances of the Rough Sets Theory*, Norwell, MA: Kluwer Academic Publishers, 1992.

ARBEE L. P. CHEN
YI-HUNG WU
Department of Computer Science, National Chengchi University, Taipei, Taiwan, R.O.C.
Department of Information and Computer Engineering, Chung Yuan Christian University, Chungli, Taiwan, R.O.C.