Establishing and retrieving domain knowledge from semi-structural corpora

*Hsien-chang WANG, +Pei-chin YANG and *Chen-chieh LI
*Chang Jung Christian University, +National Cheng Kung University
Taiwan R.O.C.

1. Introduction

1.1 Knowledge representation

The most essential part of building an expert system is the acquirement and representation of domain knowledge. In the seventies, Feigenbaum indicated the important concept of knowledge engineering. He emphasized that to utilize knowledge in problem-solving process is equally important with knowing how to solve a problem. Knowledge, according to how it is stored, can be classified to tacit knowledge and explicit knowledge. The tacit knowledge, existing in the brain of experts, can only be acquired through interviewing the domain experts. On the other hand, the explicit knowledge can be expressed clearly. Since explicit knowledge is easier to be handled, it was used in most expert systems.

Knowledge representation affects how problems are solved. Human knowledge can be expressed in the form of mathematic formulas, speech, text and figures. In artificial intelligence domain, especially in expert system research, several knowledge representation forms had been proposed (Negnevitsky, 2002). They are:

1. Semantic networks (Quillian, 1965, 1968):
   Using directed graph to represent knwolege objects and their relationship. Each object in the network is linked to other objects by their semantic relationships.
2. Case-based format (Watson, 1997; Kolodner 1993):
   Knowledge is stored in the form of cases-solutions.
3. Rule-based format (Triantaphyllou & Felici, 2006):
   If-Then rules are stored as the knowledge source.
4. Frame-based format (Minsky, 1975):
   Objects are divided into several frames, and each frame contains its corresponding attribute to describe the characteristics of the objects.
5. Ontology (Munn, 2009 ; Uschold & Gruninger, 1996):
   It is a representation of some pre-existing domain of reality which reflects the properties of the objects within its domain in such a way that there obtains a systematic correlation between reality and the representation itself. It is formalized in a way that allows it to support automatic information processing.

In this study, we adopted the concept of both ontology and frame-based approach for the knowledge representation.

1.2 Domain Knowledge (Eco-knowledge)

Many researches in 70's revealed that attempting to make a general purpose intelligent system is an unrealistic idea (Newell and Simon, 1972). The inference engine for a general purpose system is hard to build, the knowledge base is also difficult to accumulate and integrate. To make intelligent systems feasible for real applications, it was suggested that one should focus his application on a specific domain. Thus, domain knowledge plays a very important role in the realization of an intelligent system.

Accompany with the raising of eco-consciousness, going outdoor for an eco-tourism becomes a popular activity in these days. During an eco-tourism, people observe many animal and plant species, and will like to know about their names, characteristics, behaviors and further knowledge. To acquire eco-knowledge, people can listen to the explanation of a narrator or consult illustrated handbooks. However, it would be wonderful if we can build an intelligent system which is able to answer queries about specific eco-knowledge.

This goal of building an intelligent eco-knowledge system engenders three problems: (1) it requires large amount of labor to sort out all the domain knowledge; (2) it has to deal with the problem that sentences with different wording maybe describe the same fact; (3) non-expert person may not familiar with those proper nouns used by the narrator and the handbooks.

Thanks to the massive progress of linguistic processing techniques, it is possible to deal with large amount of corpora to extract the most meaningful part for flexible applications. Also, the pattern matching techniques enable the matching and discrimination between different terms more efficiently and correctly. Thus it is very possible to make our goal realize, i.e., to build an intelligent system for ordinary people to inquire eco-knowledge.

In Taiwan, the fact that there are over 500 wild bird species and eco-tourism being more popular makes bird watching a prevalent activity. Thus, this study is aimed to build an intelligent system for wild bird knowledge inquiring. The specific aims are: (1) to define the major key-features of the wild birds; (2) to collect the descriptions of wild birds; (3) to extract the linguistic features of the corpora; (4) to build up structural domain expertise automatically; (5) to define a coding schema for transforming key-features into lexical vectors; (6) to define the membership values of key-features; (7) to evaluate the similarity measurement of different key-features; (8) to illustrate the top-N answers for the input inquires.

2. Materials and Methods

Our research framework consists of four major parts as shown in Figure 1. They are described in detail in the following paragraphs.

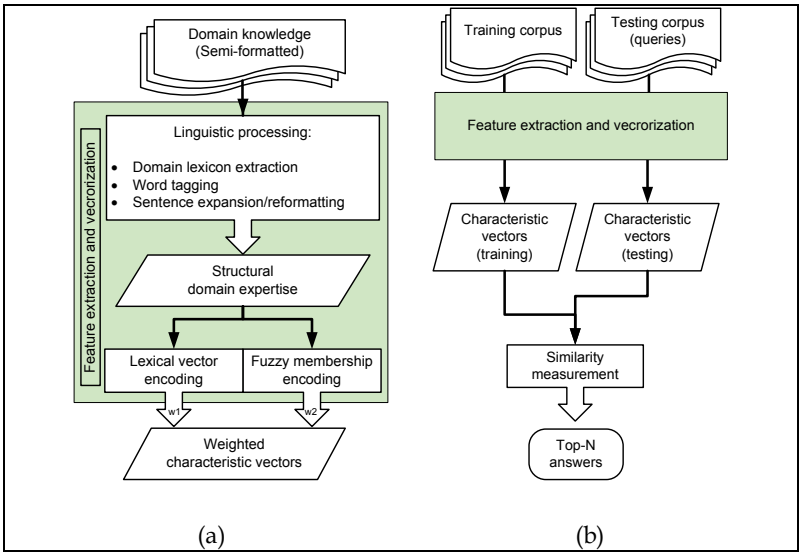

Fig. 1. The research framework for the intelligent wild bird knowledge system. (a). the training phase; (b). the testing phase.

2.1 Semi-structural Corpora

The input corpora are the so-called semi-structural domain knowledge, which contained the descriptions of each of the 442 wild bird species in Taiwan. Here, semi-structural means that the descriptions in the corpora seem follow a certain structural, however, variations often can be observed in such corpora. For example, the following sentences are all describing a fact that the bird has a white tail:

- It has a white tail.
- Its tail is white.
- It has a tail in white.

The above sentences, although in slightly different format, all consist of the part (of bird) and the attributes. Actually, it is the common format for most eco-knowledge descriptions found in the illustrated books. When describing an object (specie), the sentences are represented in the following form:

object	->	{[part]}
[part]	->	[part name] + {[attributes]}
[part]	->	{[attributes]} + [part_name]
[attribute]	->	[color] + [texture] + [shape] + [modifier]
[part_name]	->	{head, neck, tail, wing, back, beak …}
[color]	->	{black, brick-red, red, brown, dark grey …}
[modifier]	->	{long, shinny, tiny, conspicuous …}

Words in braces may appear repeatedly; words in square brackets are variable terms which can be further decomposed into other components; words without brackets are final

symbols, i.e., those which found in the original descriptions. The next step is to define the final symbols, i.e., the domain lexicon to reduce the complexity of processing.

2.2 Linguistic Processing

As shown in the previous section, the domain lexicon contains five types: [part_name], [color], [texture], [shape] and [modifier]. Since [part_name] contains the domain specific words, it has to be defined first by domain experts. The other four types of lexicon can be derived automatically by applying linguistic processing tools: CKIP AutoTag (CKIP, 2009) and HowNet (Dong & Dong, 2009).

In order to derive the domain lexicon, we apply the auto-tagging program, which is developed by CKIP group, to perform word segmentation and obtain the POS (part-of-speech) tags of each word. The semi-structural corpora are fed to the auto-tagging program, and the resulting POS-tagged words are then processed by HowNet (Dong & Dong, 2009).

HowNet is an on-line common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts. For each meaningful word, HowNet provide its semantic attributes. For instance, we can extract words with the attribute "color" easily from HonNet. Thus, by examining all the processed words, we can group those words with attribute "color" together and thus form the [color] domain lexicon. Same procedure can be applied to the [part], [texture] and [modifier] domain lexicons.

The derived domain lexicon may contaian words which are too rare or too detailed. It will cause further processing inefficient. Thus, those lexicon need to be refined. For the four types of lexicon, i.e., [part], [color], [texture] and [shape], the refinement processing is described below.

The [part] lexicon originally contain more than 130 words, however, they can be reduced to the most foundamental parts. For example, the words {forehead, upper head, backhead, hair, tophead} are reduced to the foundamental form „head".

The [color] lexicon contains 152 color words. Based on the theory of basic color (Berlin & Kay, 1969), they are reduced to 11 foundamental colors: {black, grey, white, ping, red, orange, yellow, green, blue, purple, brown}.

The [texture] lexicon is reduced to 16 words: {M-shape, Z-shapre, V-shape, fork-shape, T-shape, triangular, mackerel scale, worm hole, round, wave, point, line, thick spot, thin spot, horizontal, vertical}.

The [modifier] lexicon contains those with HowNet attributes „modifier". Those words are used as emphasized words, such as {striking, shinny, straight, interlaced, ...}

2.3 Vector Encoding

Each sentence in the corpora is transformed into two types of vectors, i.e., the ***lexical vector*** and the ***fuzzy vector***. The lexical vector concerns the lexical part of the described sentences. Lexical vector encoding is simply binary encoding. The elements of the lexical vector are either 0's or 1's. The dimension of lexical vector equals to the number of all reduced lexicon terms. (That's why we reduced the lexicon terms as mentioned above, i.e., to reduce the dimension for faster processing). For each word in the sentence to be encoded, it causes an 1 in the corresponding dimension of the vector.

The fuzzy vector of a sentence consists of the membership value between sentence-words and lexicon-words. The membership values are divided into three types: part, color and texture. We can use three tables to illustrate how fuzzy vector encoding is done.
For fuzzy membership of [part], each detailed-part word is valued by the relationship of how it closes to the fundamental parts. This process is done by averaging several expert's opinions of the membership values. An example membership table is shown below:

		Fundamental parts							
		Head	Back	Tail	Body	Wing	Ear	Beak	...
Detailed parts	Forehead	0.8	0	0	0	0	0	0	...
	Upper beak	0.1	0	0	0	0	0	0.9	...
	...	...	...	...	...	...	...	...	...

Table 1. Membership table for detailed parts.

The [color] membership is obtained by the subjective opinions of 10 tagers. For each detailed-color word, the membership value for the eleven fundamental colors are taged and averaged to produce a membership table. An example membership table is shown below:

		Fundamental colors							
		Black	Red	White	Orange	Pink	Grey	Brown	...
Detailed colors	Dark brown	0.1	0.4	0	0	0	0	0.9	...
	Rust	0.2	0.5	0	0	0	0.1	0.6	...
	...	...	...	...	...	...	...	...	...

Table 2. Membership table for detailed colors.

The [texture] membership table can be derived by a similar process. Combining all three membership values, a sentence can then be encoded into a fuzzy vector.

2.4 Vector Similarity Measure

In vector space, the similarity of two vectors X and Y can be calculated using five methods (Manning & Schutze, 1999) as shown in the Table 3.

Similarity measure	Definition
Mathcing coefficient	$X \cap Y$
Dice coefficient	$\frac{2\lvert X \cap Y\rvert}{\lvert X\rvert + \lvert Y\rvert}$
Jaccard (or Tanimoto) coefficient	$\frac{X \cap Y}{X \cup Y}$
Overlap coefficient	$\frac{\lvert X \cap Y\rvert}{\min(\lvert X\rvert, \lvert Y\rvert)}$
Cosine measure	$\frac{2\lvert X \cap Y\rvert}{\sqrt{\lvert X\rvert \times \lvert Y\rvert}}$

Table 3. Definitions of Vector similarity.

While calculating the similarity of two vectors, most approached used the cosine measure of vector intersection angle. However, since it's hard to predict the fuzzy degree of object description made by user, fuzzy encoding vectors should not use the same similarity measure as literal vectors did. In this study, we use Cosine similarity measure for lexical vectors and Overlap coefficient for fuzzy vectors. The final similarity of two descriptions is evaluated according to the measure combining the weight of lexical similarity (S_{Lex}) and fuzzy similarity (S_{Fuz}). The similarity can be evaluated by the following formula:

$$S = \alpha \cdot S_{lex} + (1 - \alpha) \cdot S_{fuz} \tag{1}$$

The literal similarity of two vectors can be defined by equation (2).

$$S_{lex}(X, Y) = \frac{\vec{X} \cdot \vec{Y}}{|\vec{X}||\vec{Y}|} = \frac{\sum_{i=1}^{m} X_i Y_i}{\sqrt{\sum_{i=1}^{m} X_i^2} \times \sqrt{\sum_{i=1}^{m} Y_i^2}} \tag{2}$$

Where, m is the dimension of the literal vector.

Let S and T be two-dimensional matrix (table) of two sentences. The fuzzy vector similarity can be expressed as equation (3) below.

$$S_{fuz}(\vec{S}, \vec{T}) = \sum_{i=1}^{s} \max\left(olp(\vec{S}, \vec{T})\right) \tag{3}$$

Where, olp(A,B) represent the overlap coefficient of vector A, B. The equation for olp(A,B) is shown below:

$$olp(\vec{A}, \vec{B}) = \frac{\vec{A} \cap \vec{B}}{\min\left(|\vec{A}|, |\vec{B}|\right)} = \frac{|\vec{C}|}{\min\left(|\vec{A}|, |\vec{B}|\right)} \tag{4}$$

where, $C=(c_1,c_2,\ldots c_n)$, $c_i=\min(A_i,B_i)$

3. Results and Discussion

The training corpus is a popular illustrated handbook (Wang et al, 1991) with detail descriptions of the features of 442 wild birds in Taiwan. The content is highly recommended by the bird watchers in Taiwan. The structures of the descriptions are very similar, however, the sentences may not grammatically valid due to the need of reducing the page amount. There are totally 6257 sentences in the training corpus.

The testing material varies from three scopes: 1) content of 40 birds from another illustrated handbook (Wu and Hsu, 1995); 2) descriptions of 20 random chosen birds made by a domain expert; 3) naive people's descriptions of 20 randomly chosen birds. The testing handbook contains similar description format as the training one, but was published by different group of people. The expert is a senior birdwatcher with experience of bird watching more than eight years. The naive people had no experience or expertise of wild birds. Figure 2 shows the four types of description for a bird named Black-browed Barbet.

五色鳥 (Black-browed Barbet)		
Corpus source	**Description in Chinese**	**Description in English**
A	嘴粗厚，黑色，腳鉛灰色。頭部大致為藍色，額、喉黃色，眉斑雜有黑色羽毛，眼先有紅色斑點，前頸亦有紅斑。後頸、背部鮮綠色，胸以下鮮黃綠色。	Beak is thick, black, foot is lead-grey. Head is almost blue, forehead and throat is yellow, eyebrow contain black feather, red dot in front of eye, fore_neck has red dot too. Back_neck and back is bright green, yellow-green below chest.
B	頭部由鮮豔的紅、黃、青、綠、黑組成，所以才稱為五色鳥。頭部大致為藍色，額、喉黃色，眉斑雜有黑色羽毛，眼先有紅色斑點，前頸亦有紅斑。後頸、背部均為鮮綠色，胸以下鮮黃綠色。	Head consist of five bright colors: red, yellow, blue, green and black, that's why it is named "five-color bird". Head is almost blue, forehead and throat is yellow, eyebrow contain black feather, red dot in front of eye, fore_neck has red dot too. Back_neck and back is bright green, yellow-green below chest.
C	全身綠色。嘴基粗厚，鐵灰色。腳灰綠色。頭藍色，額頭跟喉黃色。眼先有紅點，眉斑黑色。	The whole body is green. Beak-base is thick, iron-grey. Foot is grey-green. Head is blue, forehead and throat is yellow. A red dot before its eye, eyebrow is black.
D	頭有紅色、黃色、藍色、綠色、黑色。頭部藍色，額頭、喉嚨黃色，頸部有紅斑。後頸部、背面綠色。	Its head is red, yellow, blue, green and black. Head is blue, forehead and throat is yellow, neck contain red dot. Back_neck and back is green.

Table 4. Example of descriptions for the Black-browed Barbet. The description comes from the (A) training handbook, (B) testing handbook, (C) expert and (D) naive people, respectively.

The experiments were performed with the weight factor α set to 0, 0.1, 0.3, 0.5, 0.7, 0.9 and 1 respectively. The value of α is equal to zero if we want to ignore the lexical score; and is equal to one if we want to ignore the fuzzy score. For each testing bird, the top-N scores are recorded with N=1, 3, 5, and 10. The experimental results are shown in Figure 3 and Figure 4.

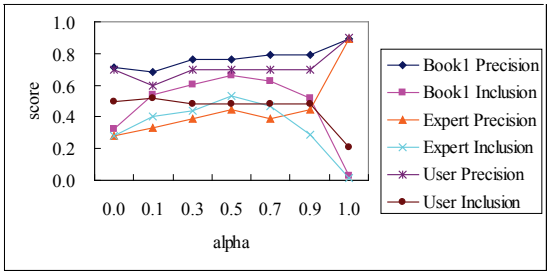

Fig. 3. The precision and inclusion rates for handbook, expert and naive user with alpha ranged from 0 to 1.

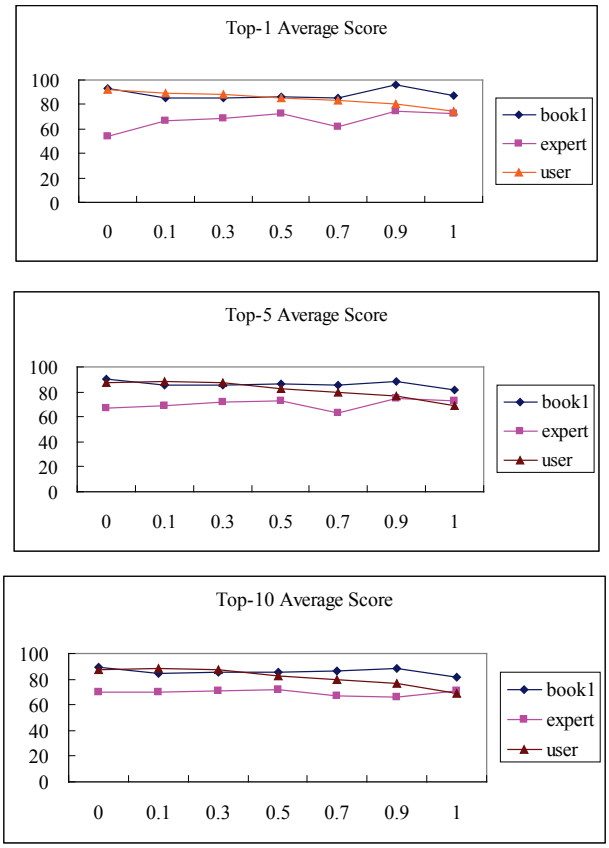

Fig. 4. The averaged score of top-N results for handbook, expert and naive user with alpha ranged from 0 to 1.

The first experiment is to compare the precision and inclusion rate of different testing data. Suppose the number of total testing data set is K, the number of correct answers appeared in the top-N candidates is C. The precision rate is defined as:

$$\text{Precision rate} = \frac{C}{K} \tag{5}$$

Suppose the total number of top-N candidates is T, the inclusion rate is defined as:

$$\text{Inclusion rate} = \frac{C}{T} \tag{6}$$

The precision rate, as defined in usual cases, tells if the correct answer is retrieved. The inclusion rate shows whether redundant answers are also reported while retrieving the answers. Figure 3 shows that, if the weighting (α) of lexical vector closed to 1.0, the precision

rate will be high and, the inclusion rate will be low. This is because redundant answers with the same similarity scores are also retrieved if consider only the lexical scores.
In Figure 4, the average matching score of expert increases as the value of α moving from 0 to 1. This is because that the wording of expert is similar to those in the handbook and thus has higher score while using large α value. (Note that higher α means higher lexical weighting.)
On the other hand, the score of naive is higher when choosing smaller α value. That is, the weighting of fuzzy vector affects the similarity score. This result corresponds with the fact that naive people are not familiar with the domain specific wordings, and introducing the fuzzy vector score has the advantage of compensating the mismatch between their wordings to those in the training corpus.
Fig. 5 and Fig. 6 show the precision rate and inclusion rate of top-10 results for all three types of testing corpora. The notation in these two figures are:

- B2: corpos from another illustrated handbook;
- E: corpos from a domain expert;
- U: corpos from a noive user;
- _P: precesion ratio;
- _I: inclusion ratio;
- _1: use cosine measure for literal and fuzzy similarity;
- _2: use overlap measure for literal and fuzzy similarity;
- _3: use cosine for literal similarity and overlap as fuzzy similarity;

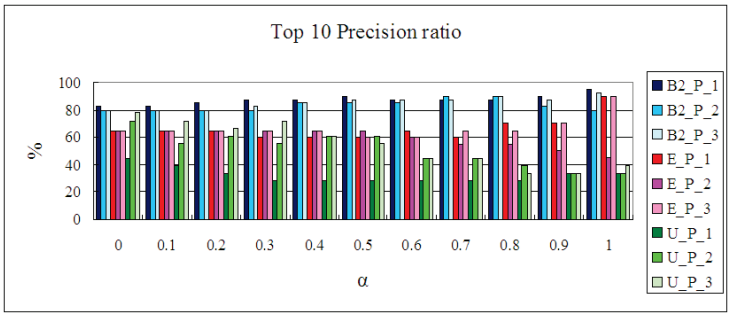

Fig. 5. Top-10 precision ratio for all testing corpora.

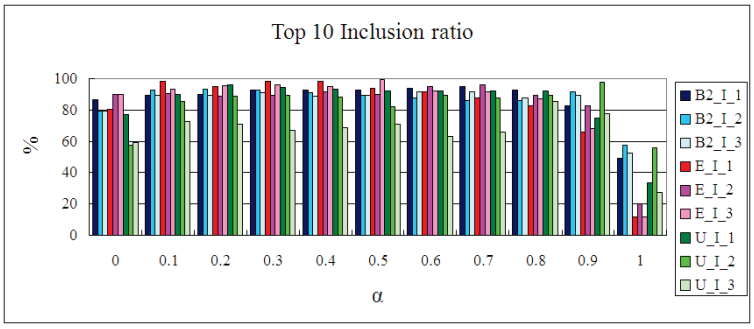

Fig. 6. Top-10 inclusion ratio for all testing corpora.

Since the training corpus is the descriptions of wild bird in an illustrated handbook, corpus B2 (another illustrated book) had the best average precision ratio; corpus E (domain expert) also achieve good result; however, corpus U (naive user) can only got good result when α is closed to 0. The results showed that to allow user query in spontaneous descripiton, the system should have high weighting in the fuzzy vector instead of literal vector.

4. Conclusion and Future Works

In this study, we proposed an approach to establish and retrieve domain knowledge automatically. The domain knowledge is established by combining the method of linguistic processing and frame-based representation. The features of descriptions consist of two major types: literal vectors and fuzzy vectors. The cosine and overlap measure is chosen to compute the similarity between literal vectors and fuzzy vectors respectively.

According to our study, several results were observed:

1. The proposed approach for domain knowledge processing is useful for establishing and retrieving eco-knowledge.
2. For some birds, its features maybe marked directly on the figures in the book, a few descriptions may be missed in the text data. This will cause some mismatch in the experiment.
3. If an experienced bird watcher wants to use the inquiry system, the literal weighting should be increased. Experiment results showed that the weighting factor could be set as 0.9.
4. For a naive use to user the inquiry system, the literal weighting should be decreased. The weighting factor could be set as 0.2.

For queries made by expert, it seems that only lexical matching is enough. However, for naive people who have no expertise on how to use specialized wording for the description of birds, combining lexical vector score with the fuzzy ones is a good choice.
Since color attributes are essential for discrimination of birds, it plays an important role in the visual cognition of birds. Currently, our study adopted only the eleven basic colors, more sophisticate color membership determination should be considered to obtain better results.
The further interesting research topic will be discovering the commonality and difference between book-style knowledge and knowledge collected from large amount of spontaneous description about objects.

5. References

Berlin B. & Kay P. (1969). *Basi Color Terms : Their Universality and Evolution,* University of California Press.
CKIP, (2009), CKIP AutoTag, available at http://ckipsvr.iis.sinica.edu.tw/
Dong Z. & Dong Q. (2009). *HowNet Knowledge Database*, http://www.keenage.com.
Kolodner J. (1993). *Case-Based Reasoning,* Morgan Kaufmann, 978-1558602373.
Manning C. D. & Schutze H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, 978-0262133609, Cambridge.

Minsky, M. (1975). A framework for representation knowledge. *The Psychology of Computer Vision*, McGraw-Hill, 978-0070710481, New York.

Munn K. & Smith B. (2009). *Applied Ontology, An Introduction,* Ontos Verlag Transaction Pub, 978-3938793985.

Negnevitsky M., (2002). *Artificial Intelligence, A Guide to Intelligent Systems,* Addison-wesly, 978-0321204660, England.

Newell A. & Simon H.A. (1972). *Human Problem Solving*, Prentice Hall, Englewood Cliffs, 978-0134454030, NJ.

Quillian M.R. (1965). Word concepts: a theory and simulation of some basic semantic capabilities, *Behavioral Science*, Vol. 12, No. 5., pp. 410-430.

Quillian M.R. (1968). Semantic Memory, *Semantic Information Processing,* The MIT Press, Ch. 4, pp.227-270, 978-0262130448.

Triantaphyllou E. & Felici G. (2006). *Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques (Massive Computing),* Springer, 978-0387342948.

Wang G.H. et al., (1991). *Taiwan Wild Birds, Arthur Books, Taipei.*

Watson I. (1997). *Applying Case-Based Reasoning: Techniques for Enterprise Systems,* Morgan Kaufmann, 978-1558604629.

Wu T.H. & Hsw W.B. (1995), *Guiding Map of Bird Watching in Taiwan,* BigTree Culture, Taipei.

Uschold, M., Gruninger, M. (1996). *Ontologies: Principles, Methods and Applications*, The Knowledge Engineering Review, 11, 93-136.

Machine Learning

Edited by Yagang Zhang

ISBN 978-953-307-033-9
Hard cover, 438 pages
Publisher InTech
Published online 01, February, 2010
Published in print edition February, 2010

Machine learning techniques have the potential of alleviating the complexity of knowledge acquisition. This book presents today's state and development tendencies of machine learning. It is a multi-author book. Taking into account the large amount of knowledge about machine learning and practice presented in the book, it is divided into three major parts: Introduction, Machine Learning Theory and Applications. Part I focuses on the introduction to machine learning. The author also attempts to promote a new design of thinking machines and development philosophy. Considering the growing complexity and serious difficulties of information processing in machine learning, in Part II of the book, the theoretical foundations of machine learning are considered, and they mainly include self-organizing maps (SOMs), clustering, artificial neural networks, nonlinear control, fuzzy system and knowledge-based system (KBS). Part III contains selected applications of various machine learning approaches, from flight delays, network intrusion, immune system, ship design to CT and RNA target prediction. The book will be of interest to industrial engineers and scientists as well as academics who wish to pursue machine learning. The book is intended for both graduate and postgraduate students in fields such as computer science, cybernetics, system sciences, engineering, statistics, and social sciences, and as a reference for software professionals and practitioners.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821