Automatic Bibliographic Component Extraction Using Conditional Random Fields

Sheng-Ming Wang[1], Wei-Pang Yang[2], Hsin-Ping Chou[2], Fu-Mei Chen[3], Jia-Li Hou[2], Jyh-Jian Sheu[4]
[1]Graduate Institute of Interactive Media Design, National Taipei University of Technology, Taiwan
[2]Department of Information Management, National Dong Hwa University, Taiwan
[3]Research and Development Center, National Taipei University of Technology, Taiwan
[4]College of Communication, National Chengchi University, Taiwan
ryan5885@mail.ntut.edu.tw, wpyang@mail.ndhu.edu.tw, shean@ms49.url.com.tw
chenfmei@mail.ntut.edu.tw, alexhou@mail.ndhu.edu.tw, jjsheu@nccu.edu.tw

Abstract

Bibliographic data and publication data are composed of subfields such as "author," "title," "journal," and "year." Citation analysis of articles in scholarly journals is a very effective method for their evaluation. This paper proposes a system for analyzing bibliographic component strings, which is based on the technique of Conditional Random Fields (CRF). The system is composed of two major modules: the Bibliographic Extraction Module (BEM) and the Statistical Evaluation Module (SEM). The objective of the Bibliographic Extraction Module is to extract the bibliographic components based on the machine learning technique, and the objective of the Statistical Evaluation Module is to turn the extracted bibliographic information into a statistical report.

In this paper, we apply the CRF technique to build a probability model for dividing sequential data and giving proper tags to the components according to their characteristics. This is the framework for building the BEM to segment and label bibliographic information, identifying the author's name, journal's name, date of publication and so on. Then we employ the SEM to filter and match the intermediate representations produced by the BEM. In the end, the SEM will output the final evaluation report. Experimental results show that our system is reliable, with excellent overall efficiency.

Keywords: Conditional random field, Citation analysis, Machine learning.

1 Introduction

As electronic literature is disseminated in the open Internet environment, it is quite easy both to publish and to access an article. However, as more and more articles become widely dispersed on the Internet, it becomes a time-consuming and effort-intensive job to search a targeted article. It would be efficient for a researcher to analyze the citations and appraisals of articles if those articles could be arranged in a logical sequence. It would also be very convenient to offer the linkage for a citation among authors when users navigate electronic literature.

Chieu and Ng [5] define information extraction as a problem of classification. Automatic information abridgement is the process of extracting the most important information automatically from a set of data by using a computer. Some sorts of classification methods, such as regular expressions, rule-based parsers and machine learning, have been utilized on text classification [10]. Among these methods, machine learning is as suitable as symbolic learning [25], grammar induction [4], Support Vector Machines (SVM) [9], Hidden Markov Models (HMM) [26] and statistical methods [23] for automatic metadata extraction.

This research tries to extract important information such as author, article title, journal name and publication date from the sequential data of bibliographic references. For example, consider the reference "C. C. Lee, M. S. Hwang, and W. P. Yang, 'Extension of Authentication Protocol for GSM,' *IEEE Proc. Commun.*, vol. 150, no. 2, pp. 91-95, 2003." In this paper, we propose an automatic bibliographic component extraction system capable of properly dissecting important data from the author names: "C. C. Lee, M. S. Hwang, and W. P. Yang", and the article title, "Extension of Authentication Protocol for GSM." The purpose is to convert unstructured bibliographic data elements into structured ones, setting their meanings and observing them from different dimensions, thereby making it possible to classify them and search them.

The primary prerequisite of our system design is to exhibit maximal tolerance for the different bibliographic formats that result from various systems of writing. We extract each field of bibliographic components by applying the machine learning technique. According to our review, the CRF [11] technique has been utilized in such tasks as name entity extraction [14], table extraction [17] and shallow parsing [21]. CRF has the advantages of both the finite-state Hidden Markov Models (HMM) and Support Vector Machines (SVM) techniques, such as considering dependent features through sequencing. Thus, we use CRF in this research to build a probability model for dividing

*Corresponding author: Jyh-Jian Sheu; E-mail: jjsheu@nccu.edu.tw sequential data and giving the individual components proper tags according to their characteristics.

2 Literature Review

The machine learning technique has been widely used in information extraction. There are two categories of learning methods. One of them, rule-based learning, is used in the Rapier, BWI and $(LP)^2$ models, which extract required information via the training and learning that occurs from following given rules. The other category of learning methods is statistical machine learning, as exemplified in Maximum Entropy, HMM and SVM.

In this section we present and compare three machine learning methods frequently used in information extraction. The Support Vector Machines (SVM), Hidden Markov Models (HMM), and Conditional Random Fields (CRF) models are reviewed.

2.1 Support Vector Machines

Support Vector Machines (SVM) is a learning tool [24] developed by Vapnik and his co-workers for data classification, regression and pattern recognition. The learning method of statistical learning theory, based on the Kernel Function, has many unique advantages in solving small-sample, non-linear and high-dimensional pattern recognition issues and has gained outstanding results on pattern recognition, function approximation and probability density estimation.

The applications of SVM are various, such as text categorization, image recognition [2], hand-written digit recognition [1], data mining and bioinformatics [8]. SVM is built upon the concepts of statistical learning, neural networks and optimal theory. It features (1) the ability to handle both linear and non-linear problems, and (2) no limitation of data volume. Therefore, SVM algorithms can provide an effective means of solving the categorization problems of high-volume data. SVM uses the binary classification method to sort the data [15] and can process many dependent features to map N-dimensional input space into a high-dimensional feature space with a non-linear classifier.

Basically, SVM is a kind of forward neural network in nature. According to the inference of structured risk minimization, and under the assumption of minimizing the error of training samples, we have raised the generalized capability of the classifier as high as possible in our research. From the implementation point of view, the core concept of training SVM is equivalent to solving a linear constrained quadratic planning problem, thus constructing a hyper plane as a decision platform that maximizes the distance between two classes in the characteristic space and guarantees the best global solution.

2.2 Hidden Markov Model

The Hidden Markov Model (HMM) was first described in a series of statistical theses by Leonard E. Baum and other scholars in the late 1960s. One of its first applications was voice recognition, commencing in 1989 [18]. In addition to voice recognition, HMM is now also applied to optical signal recognition, machine translation, bioinformatics and genome analysis.

According to the definition given by Lawrence in 1989 [12], the HMM is a statistical model which is used to describe the Markov process implied with unknown parameters. Its challenge is to determine the implied parameters from observable ones and thus to make further analysis with them. The states of the HMM are uncertain or invisible. The observed events and the states are not in one-to-one correspondence, though they are correspondingly related through a set of probability distribution. The HMM utilizes a dual-random process composed of two parts; the first part consists of Markov Chains, which describe the transition of the states, and the other part is the general random process that describes the relationship between the states and the series of observations.

Some researchers have previously applied the HMM technique to information retrieval. In 1997, Bikel used the HMM to retrieve nouns (such as “Price”) from unstructured documents [3]. In 2000, Freitag retrieved related phrases from documents with unrelated words [7]. Leek used the HMM to retrieve the information of locations and gene-related nouns from documents [13].

In order to define the HMM precisely, some required elements have to be understood. Here, we use the 5-element model $\lambda = (N, M, \pi, A, B)$ to describe the HMM. Please refer to Table 1 below:

Table 1 HMM Basic Elements

Parameter	Meaning
N	No. of states
M	No. of possible observations of each state
A	Matrix of time-independent state transition probability
B	Observation probability distribution under given state
π	Space probability distribution of initial state

2.3 Conditional Random Fields

Conditional Random Fields (CRF) is a probability model for dividing and marking structured data as sequences, matrices and trees [11]. Lafferty mentioned that CRF is trained by the indirect graphic model to maximize the conditional probability [11]. Through sequentially tagging specific parts of the observation to define the conditional probability distribution, CRF is superior to the HMM regarding the nature of the conditions. Besides, CRF also avoids the symptom of tagging deviation. Currently, CRF techniques are used in fields such as named entity recognition [14], table extraction [17], shallow parsing [21] and flexible features learning [22].

In the further development of the CRF tool, Kudo has proposed a CRF++ tool for information extraction [13]. CRF++ is an open-source toolset which can construct a simple way to solve problems according to specified requirements. It can be used to classify the elements from sequential data, corresponding to their tags. It is designed to fulfill the general requirements for Natural Language Processing (NLP) tasks, such as named entity recognition, information extraction and text dividing.

A prerequisite to using the CRF++ tool is knowledge of the format of the training and testing files. The format of the training and testing files has to contain multiple tokens with multiple fields. Each token must either be written on a single line and separated by a space or be put into a table. The sequence of the tokens can form a sentence. Sentences are separated by a space line. Each token is depicted as three fields:

(1) Word itself, such as “reckons.”
(2) Part-Of-Speech (POS), such as “VBZ.”
(3) Divided Tag in IOB2 format.

As for the Part-Of-Speech (POS) function, the NLP Processor, developed by the University of Edinburgh [14], uses the enhanced Penn Treebank Tag-set for training. The training set is around 1 million words compiled from documents on the Internet. The NLP processor determines the grammatical classification of each word of a sentence.

Here, we use the following sentence as an example to explain the training and testing files: “*He reckons the current account deficit will narrow to only #1.8 billion in September*.” The format of these files should be transformed into the three categories of Word, POS and Tag for each token. Therefore, the example input value is transformed into the format shown in Table 2.

Table 2 Example of the Training and Testing File Format

Word	POS	Tag
He	PRP	B-NP
Reckons	VBZ	B-VP
the	DT	B-NP
Current	JJ	I-NP
Account	NN	I-NP
Deficit	NN	I-NP
Will	MD	B-VP
Narrow	VB	I-VP
To	TO	B-PP
Only	RB	B-NP
#	#	I-NP
1.8	CD	I-NP
Billion	CD	I-NP
In	IN	B-PP
September	NNP	B-NP
.	.	0

In the training stage, we use the Crf_learn command as follows:

% Crf_learn template_file train_file model_file

Note that the template_file and the train_file should be prepared by the user in advance. The Crf_learn command will create the training model and store it in the model_file.

In the testing stage, we use the Crf_test command as follows:

% Crf_test -- m model_file, testing_files …

The model_file was built by the Crf_learn command; thus, we do not need to build it. In the process of testing, users need not specify the template_file, as the model_file already has the template information. The testing_file contains the testing material you will use for marking the tags. Table 3 is an example of the result of a Crf_test.

Table 3 Example of Crf_Test Output

Word	POS	Tag	Prediction
Rockwell	NNP	B	B
International	NNP	I	I
Corp.	NNP	I	I
‘s	POS	B	B
Tulsa	NNP	I	I
unit	NN	I	I
..			

The last column contains the prediction tags. The accuracy rate can be calculated by simply comparing the differences between the third and fourth columns, if column three contains the standard tag.

3 Automatic Bibliographic Extraction Based on CRF

We use CRF to solve the problems of bibliographic component extraction by turning unstructured sequential data into structured ones by trying to extract the components correctly, and giving each component a proper tag and meaning according to its feature.

In this section we will first define the problems of bibliographic component extraction by investigating input and output data. This is, followed by an introduction to the operation of each module in the system and then a discussion of the design and construction of the whole system. System implementation results will be shown in the final part of this section.

3.1 Definition of Problems

The problems of bibliographic component extraction are similar to those of information extraction. We need to extract required information, such as author, year, journal title etc. However, each article may use a specific article format, in which the writing of the name or the acronyms is different. Dealing with the various possible formats of a single article is the key issue in handling bibliographic component extraction. As an example, the following bibliographic entries reference the same book but with different formatting.

Format 1: L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. Classification and Regression Trees. Wadsworth, Pacific Grove, California, 1984.

Format 2: L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth and Brooks, 1984.

Format 3: L. Breiman et al. Classification and Regression Trees. Wadsworth, 1984.

3.2 System Operation Flow

The bibliographic component extraction system proposed in this research, which uses CRF as a background technique, is composed of two major modules: the Bibliographic Extraction Module (BEM) and the Statistical Evaluation Module (SEM).

The objective of the BEM is to extract the bibliographic components based on the machine learning technique. The objective of the SEM is to turn the extracted bibliographic information into a statistical report. As noted above, the formatting of the bibliographic components may not be uniform.

Figure 1 shows the operational flow of the BEM. This module is further divided into a Training Phase and a Testing Phase. In the Training Phase, we enter the Feature Template (containing the features used in training and testing) and the Training Corpus (containing training data, which were tagged manually in advance). The system with learning capability extracts the representational features automatically from the Training Corpus and uses its corresponding summary to produce the rule and construct the trained model by using CRF learning algorithms.

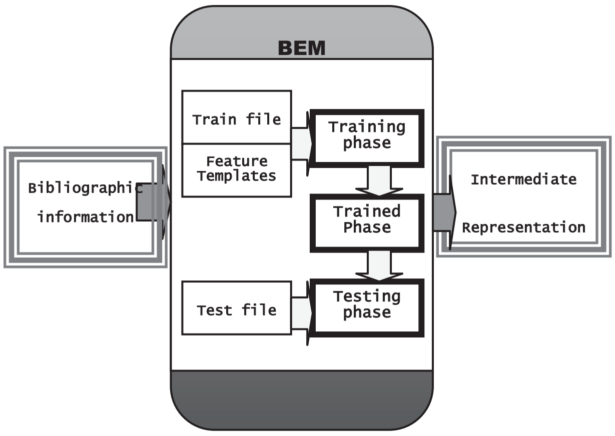


Figure 1 Bibliographic Extraction Module

In the Testing Phase, enter the Testing Corpus (testing data) with the same format as, but not belonging to, the Training Corpus. The system will extract representational features by learned rules that combine with the trained model to produce the estimated summary of the Testing Corpus.

For example, in the Training Phase, we enter the Training Corpus with manually entered tags, such as “<name>Hsin-Ping Chou</name>” and “<name>LBLin</name>,” to produce the training model. When entering “CCLee” during the Testing Phase, the system will produce the estimated summary of the Testing Corpus by learned rules and the output is the intermediate representation “<name>CCLee</name>.”

After receiving the intermediate representation, the SEM calculates and filters the data by the conditions of author, title, journal, and year and then summarizes the bibliographic components that meet the specified year and journal category conditions to generate the article evaluation report, as shown in Figure 2.

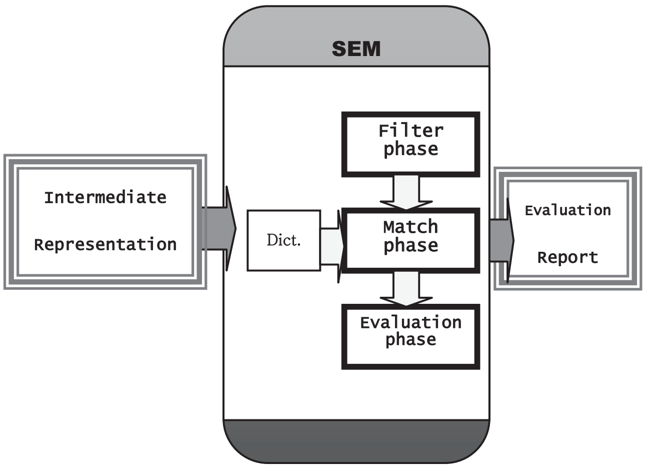


Figure 2 Statistical Evaluation Module

As an example, suppose there is a piece of intermediate representation as follows:

<name>CCLee</name>
<title> Extension of Authentication Protocol for GSM</title>
<journal> IEEE Procedure Communication </journal>
<year>2003</year>

Suppose that the journal articles of CCLee between the year of 2001 and 2005 should be classified into several categories (say, SCI and SSCI) to count the total number. When the above intermediate representation enters the SEM module, the SEM will begin to check if the author is CCLee, and then check if the publication date is between 2001 and 2005. In the second step, the module will check to determine whether the journal belongs to SCI or SSCI. The intermediate representations of other articles will be processed in the same way. Finally, the SEM summarizes and calculates the total number of articles that meet the specified conditions and generates the article evaluation report for CCLee.

In summary, the system operation flow begins with entering the bibliographic information to the BEM to generate the intermediate representation through the Training Phase. The intermediate representation is then processed by the SEM to generate the final article evaluation report.

3.3 System Design

In the Training Phase, we first enter the manually tagged Training Corpus and Feature Template. Note that the format of the Training Corpus and Testing Corpus are the same, although they do not contain the same data set. The corpus has to contain multiple tokens, with each word treated as a token. As a result, every token contains three fields:

(1) Word, such as Hsin-Ping.
(2) Part-Of-Speech (POS), such as NNP.
(3) Divided Tag shown in IOB2 format, such as B-PERSON.

Every token and its POS and tag should be written on one line. A series of tokens can compose a sentence. Every row is separated by a space or with the table. Every sentence is separated by a space line.

For example, suppose the input value of the Training Corpus and Testing Corpus is:

C. C. Lee, "Extension of Authentication Protocol for GSM," IEEE Procedure Communication, vol. 150, no. 2, pp. 91-95, 2003.

The input value is first transformed into the 3-column format of the corpus (i.e., word, POS, tag), as shown in Table 4:

Table 4 Format of Training Corpus and Testing Corpus

Word	POS	Tag
C	NNP	B-PERSON
.	.	0
C	NNP	I-PERSON
.	.	0
Lee	NNP	I-PERSON
,	,	0
“	“	0
Extension	NN	B-TITLE
Of	IN	I-TITLE
Authentication	NN	I-TITLE
Protocol	NN	I-TITLE
For	IN	I-TITLE
GSM	NNP	I-TITLE
,	,	0
“	“	0
IEEE	NNP	B-ORGANIZATION
Procedure	NN	I-ORGANIZATION
communication	NN	I-ORGANIZATION
,	,	0
Vol	NNP	0
.	.	0
150	CD	0
,	,	0
pp	NNP	0
.	.	0
91	CD	0
-	-	0
95	CD	0
,	,	0
2003	CD	B-YEAR
.	.	0
..	..	0

Second, we prepare the Feature Template, which contains the features used in both the training and testing phases. Its basic format is %X [row, col]. It is used to determine a token in the input data, where "row" specifies the row number relative to the current row and "col" is the absolute column number. Table 5 shows an example of the Feature Template that uses the third row as the current row:

If it is necessary to clarify the relative position of a token, the user can use a mark for this purpose. For example, in Table 6, both "%x[0,1]" and "%x[-2,1]" represent "NNP". However, they are distinct ones. In

Table 5 Example of the Feature Template

Training Corpus	col0	col1	col2
r-2	C	NNP	B-PERSON
r-1	.	.	0
r0	C	NNP	I-PERSON
r1	.	.	0
r2	LEE	NNP	I-PERSON

Table 6 Feature Template

Template	Extended Feature
%x[0,0]	C
%x[0,1]	NNP
%x[-1,0]	.
%x[-2,1]	NNP
%x[0,0] / %x[0,1]	C/NNP
ABC%x[0,1]123	ABCNNP123

order to distinguish between them, an independent mark ("U01" or "U02") can be added into the template, such as "U01:%x[0,1]" and "U02:%x[-2,1]". Under such circumstances, the two templates will be treated as different ones.

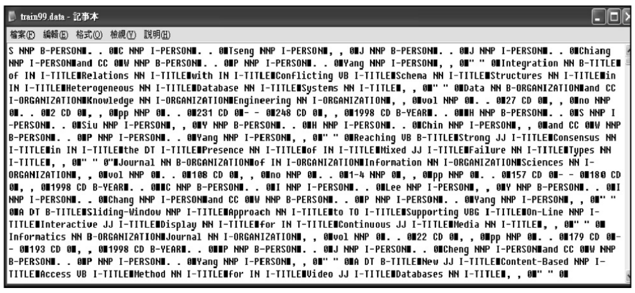
Additionally, only the letters of the author's name are extracted, while any punctuation marks are discarded. For example, "C.C.Lee" will be extracted as "CCLee." This is to avoid mistakes caused by the punctuation when processing the author extraction. Acronyms will be extracted directly and treated as proper nouns. The comparison of the acronyms is left to be treated in the SEM. In regards to dashes, or hyphens, there are two different treatments. When a dash appears with numbers, such as 181-195, the 181 and the 195 are treated as numbers while the dash is treated as a symbol. In the case of words with dashes or hyphens, "Hsin-Ping" for example, it is all treated as a single word.

3.4 The Bibliographic Extraction Module

In the BEM development, the Training Corpus, shown in Figure 3, and the Feature Template, shown in Figure 4, need to be prepared in advance. During the Training Phase, the command Crf_learn uses the method of CRF to extract the representational features automatically from the Training Corpus. Then, the extracted features are stored into the trained model, and the corresponding rules are produced at the same time.

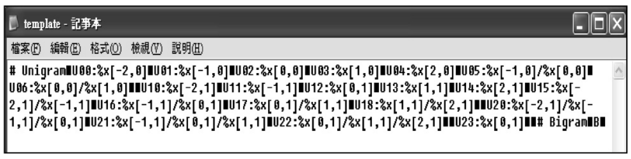
The command Crf_learn will create the trained model and store it in the model file by using the following command:

% Crf_learn template_file train_file model_file



```
S NNP B-PERSON■. . 0■C NNP I-PERSON■. . 0■Tseng NNP I-PERSON■, , 0■J NNP B-PERSON■. . 0■J NNP I-PERSON■. . 0■Chiang
NNP I-PERSON■and CC 0■W NNP B-PERSON■. . 0■P NNP I-PERSON■. . 0■Yang NNP I-PERSON■, , 0■" " 0■Integration NN B-TITLE■
of IN I-TITLE■Relations NN I-TITLE■with IN I-TITLE■Conflicting VB I-TITLE■Schema NN I-TITLE■Structures NN I-TITLE■in
IN I-TITLE■Heterogeneous NN I-TITLE■Database NN I-TITLE■Systems NN I-TITLE■, , 0■" " 0■Data NN B-ORGANIZATION■and CC
I-ORGANIZATION■Knowledge NN I-ORGANIZATION■Engineering NN I-ORGANIZATION■, , 0■vol NNP 0■. . 0■27 CD 0■, , 0■no NNP
0■. . 0■2 CD 0■, , 0■pp NNP 0■. . 0■231 CD 0■- - 0■248 CD 0■, , 0■1998 CD B-YEAR■. . 0■■H NNP B-PERSON■. . 0■S NNP I
-PERSON■. . 0■Siu NNP I-PERSON■, , 0■Y NNP B-PERSON■. . 0■H NNP I-PERSON■. . 0■Chin NNP I-PERSON■, , 0■and CC 0■W NNP
B-PERSON■. . 0■P NNP I-PERSON■. . 0■Yang NNP I-PERSON■, , 0■" " 0■Reaching VB B-TITLE■Strong JJ I-TITLE■Consensus NN
I-TITLE■in IN I-TITLE■the DT I-TITLE■Presence NN I-TITLE■of IN I-TITLE■Mixed JJ I-TITLE■Failure NN I-TITLE■Types NN
I-TITLE■, , 0■" " 0"■Journal NN B-ORGANIZATION■of IN I-ORGANIZATION■Information NN I-ORGANIZATION■Sciences NN I-
ORGANIZATION■, , 0■vol NNP 0■. . 0■108 CD 0■, , 0■no NNP 0■. . 0■1-4 NNP 0■, , 0■pp NNP 0■. . 0■157 CD 0■- - 0■180 CD
0■, , 0■1998 CD B-YEAR■. . 0■■C NNP B-PERSON■. . 0■I NNP I-PERSON■. . 0■Lee NNP I-PERSON■, , 0■Y NNP B-PERSON■. . 0■I
NNP I-PERSON■. . 0■Chang NNP I-PERSON■and CC 0■W NNP B-PERSON■. . 0■P NNP I-PERSON■. . 0■Yang NNP I-PERSON■, , 0■" "
0■A DT B-TITLE■Sliding-Window NNP I-TITLE■Approach NN I-TITLE■to TO I-TITLE■Supporting VBG I-TITLE■On-Line NNP I-
TITLE■Interactive JJ I-TITLE■Display NN I-TITLE■for IN T-TITLE■Continuous JJ I-TITLE■Media NN I-TITLE■, , 0■" " 0■
Informatics NN B-ORGANIZATION■Journal NN I-ORGANIZATION■, , 0■vol NNP 0■. . 0■22 CD 0■, , 0■pp NNP 0■. . 0■179 CD 0■-
- 0■193 CD 0■, , 0■1998 CD B-YEAR■. . 0■■P NNP B-PERSON■. . 0■J NNP I-PERSON■. . 0■Cheng NNP I-PERSON■and CC 0■W NNP
B-PERSON■. . 0■P NNP I-PERSON■. . 0■Yang NNP I-PERSON■, , 0■" " 0■A DT B-TITLE■New JJ I-TITLE■Content-Based NNP I-
TITLE■Access VB I-TITLE■Method NN I-TITLE■for IN I-TITLE■Video JJ I-TITLE■Databases NN I-TITLE■, , 0■" " 0■
```

Figure 3 Training Corpus



```
# Unigram■U00:%x[-2,0]■U01:%x[-1,0]■U02:%x[0,0]■U03:%x[1,0]■U04:%x[2,0]■U05:%x[-1,0]/%x[0,0]■
U06:%x[0,0]/%x[1,0]■■U10:%x[-2,1]■U11:%x[-1,1]■U12:%x[0,1]■U13:%x[1,1]■U14:%x[2,1]■U15:%x[-
2,1]/%x[-1,1]■U16:%x[-1,1]/%x[0,1]■U17:%x[0,1]/%x[1,1]■U18:%x[1,1]/%x[2,1]■■U20:%x[-2,1]/%x[-
1,1]/%x[0,1]■U21:%x[-1,1]/%x[0,1]/%x[1,1]■U22:%x[0,1]/%x[1,1]/%x[2,1]■■U23:%x[0,1]■■# Bigram■B■
```

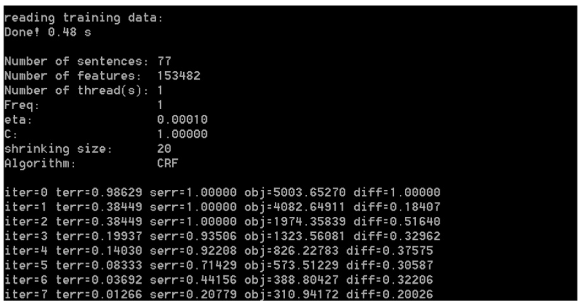
Figure 4 Feature Template

The training process that follows execution of the Crf_learn command is shown in Figure 5.

In the Testing Phase of the BEM, the system extracts the related features according to the learned rules first. Then it creates the estimated summary of the Testing Corpus by applying the learned rules and trained model (created by the Crf_learn command). In the process of testing, the user doesn't need to specify the Feature Template, as it is already in the model. Note that the format of the Testing Corpus should be the same as that of the Training Corpus. However, the contents of the Testing Corpus do not belong to the Training Corpus, and the Crf_test command can be used for testing:

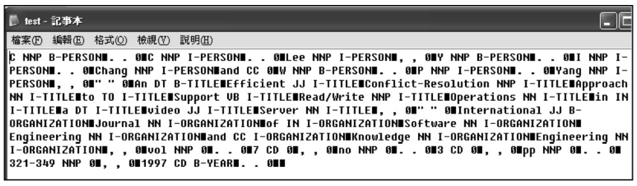
% Crf_test –m model_file,, testing_files …

as shown in Figures 6 and 7.



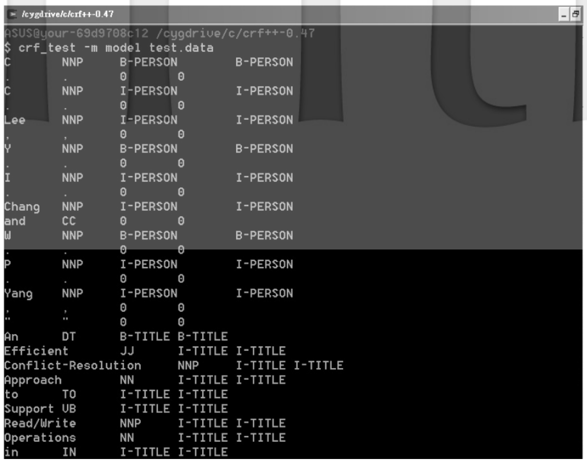
```
reading training data:
Done! 0.48 s

Number of sentences: 77
Number of features:  153482
Number of thread(s): 1
Freq:                1
eta:                 0.00010
C:                   1.00000
shrinking size:      20
Algorithm:           CRF

iter=0 terr=0.98629 serr=1.00000 obj=5003.65270 diff=1.00000
iter=1 terr=0.38449 serr=1.00000 obj=4082.64911 diff=0.18407
iter=2 terr=0.38449 serr=1.00000 obj=1974.35839 diff=0.51640
iter=3 terr=0.19937 serr=0.93506 obj=1323.56081 diff=0.32962
iter=4 terr=0.14030 serr=0.92208 obj=826.22783 diff=0.37575
iter=5 terr=0.08333 serr=0.71429 obj=573.51229 diff=0.30587
iter=6 terr=0.03692 serr=0.44156 obj=388.80427 diff=0.32206
iter=7 terr=0.01266 serr=0.20779 obj=310.94172 diff=0.20026
```

Figure 5 Training Process



```
C NNP B-PERSON■. . 0■C NNP I-PERSON■. . 0■Lee NNP I-PERSON■, , 0■Y NNP B-PERSON■. . 0■I NNP I-
PERSON■. . 0■Chang NNP I-PERSON■and CC 0■W NNP B-PERSON■. . 0■P NNP I-PERSON■. . 0■Yang NNP I-
PERSON■, , 0■" " 0■An DT B-TITLE■Efficient JJ I-TITLE■Conflict-Resolution NNP I-TITLE■Approach
NN I-TITLE■to TO I-TITLE■Support VB I-TITLE■Read/Write NNP I-TITLE■Operations NN I-TITLE■in IN
I-TITLE■a DT I-TITLE■video JJ I-TITLE■Server NN I-TITLE■, , 0■" " 0■International JJ B-
ORGANIZATION■Journal NN I-ORGANIZATION■of IN I-ORGANIZATION■Software NN I-ORGANIZATION■
Engineering NN I-ORGANIZATION■and CC I-ORGANIZATION■Knowledge NN I-ORGANIZATION■Engineering NN
I-ORGANIZATION■, , 0■vol NNP 0■. . 0■7 CD 0■, , 0■no NNP 0■. . 0■3 CD 0■, , 0■pp NNP 0■. . 0■
321-349 NNP 0■, , 0■1997 CD B-YEAR■. . 0■■
```

Figure 6 Testing Corpus



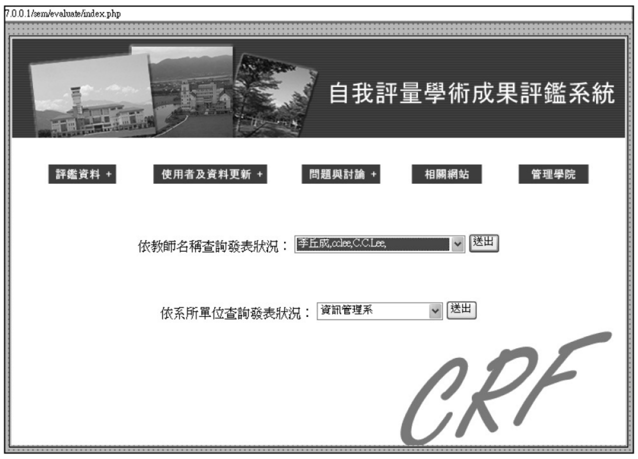
```
/cygdrive/c/crf++-0.47
ASUS@your-69d9708c12 /cygdrive/c/crf++-0.47
$ crf_test -m model test.data
C       NNP     B-PERSON        B-PERSON
.       .       0       0
C       NNP     I-PERSON        I-PERSON
.       .       0       0
Lee     NNP     I-PERSON        I-PERSON
,       ,       0       0
Y       NNP     B-PERSON        B-PERSON
.       .       0       0
I       NNP     I-PERSON        I-PERSON
.       .       0       0
Chang   NNP     I-PERSON        I-PERSON
and     CC      0       0
W       NNP     B-PERSON        B-PERSON
.       .       0       0
P       NNP     I-PERSON        I-PERSON
.       .       0       0
Yang    NNP     I-PERSON        I-PERSON
,       ,       0       0
"       "       0       0
An      DT      B-TITLE B-TITLE
Efficient       JJ      I-TITLE I-TITLE
Conflict-Resolution     NNP     I-TITLE I-TITLE
Approach        NN      I-TITLE I-TITLE
to      TO      I-TITLE I-TITLE
Support VB      I-TITLE I-TITLE
Read/Write      NNP     I-TITLE I-TITLE
Operations      NN      I-TITLE I-TITLE
in      IN      I-TITLE I-TITLE
```

Figure 7 Example of Crf_test Output

3.5 The Statistical Evaluation Module (SEM)

The intermediate representation is created after entering the article list to the BEM and completing the training and Testing Phase. The SEM will filter and compare the different conditions by year and category, and then summarize those meeting the conditions for each category respectively. Finally, the SEM will create the article classification statistics report.

Figure 8 shows an example of the inquiry of the status of C.C. Lee's articles from 2001 through 2005 in our system. Figure 9 shows the result of the statistics report, which indicates that CCLee published three articles of SCI and no articles of SSCI from 2001 through 2005.

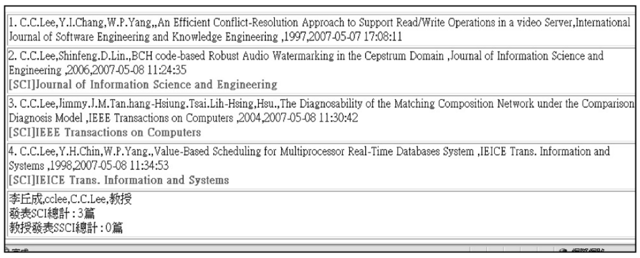


7.0.0.1/sem/evaluate/index.php

依教師名稱查詢發表狀況：李丘成,cclee,C.C.Lee,

依系所單位查詢發表狀況：資訊管理系

Figure 8 Inquiry of Professor CCLee's Issuance Status from 2001 through 2005



1. C.C.Lee,Y.I.Chang,W.P.Yang,,An Efficient Conflict-Resolution Approach to Support Read/Write Operations in a video Server,International Journal of Software Engineering and Knowledge Engineering ,1997,2007-05-07 17:08:11

2. C.C.Lee,Shinfeng.D.Lin.,BCH code-based Robust Audio Watermarking in the Cepstrum Domain ,Journal of Information Science and Engineering ,2006,2007-05-08 11:24:35
[SCI]Journal of Information Science and Engineering

3. C.C.Lee,Jimmy.J.M.Tan.hang-Hsiung.Tsai.Lih-Hsing,Hsu.,The Diagnosability of the Matching Composition Network under the Comparison Diagnosis Model ,IEEE Transactions on Computers ,2004,2007-05-08 11:30:42
[SCI]IEEE Transactions on Computers

4. C.C.Lee,Y.H.Chin,W.P.Yang.,Value-Based Scheduling for Multiprocessor Real-Time Databases System ,IEICE Trans. Information and Systems ,1998,2007-05-08 11:34:53
[SCI]IEICE Trans. Information and Systems

李丘成,cclee,C.C.Lee,教授
發表SCI總計：3篇
教授發表SSCI總計：0篇

Figure 9 Professor CCLee's Statistics of Issuance by Category

4 Experimental Results

4.1 The Experimental Data

We collected bibliographic information from the faculty database of National Dong Hwa University in Taiwan. The scope of the collection was the professors' publication lists from the Institute of Science & Engineering and the Institute of Management. Our study required two separate bibliographic collections: the Training Corpus and the Testing Corpus. To make these two collections similar yet unique, we extracted articles of the same type (category) but with different forms and contents. It's worth mentioning that we only extract information from the publication list or the reference section. This is quite different from other methods that extract from the content of articles.

The Training Corpus we compiled contains 100 articles collected from eight departments of the two institutes. Each article is characterized by four bibliographic items: Author, Title, Journal and Year. The Testing Corpus contains another 50 articles. The selection criteria of each corpus are as different as possible on the bibliographic items and the composition sequence of items. We used these corpuses to test the bibliographic extraction performance of our system for various items and sequences. Although the type of information collected in the Training Corpus and Testing Corpus is the same, the fact that they contain different bibliographic data maintains the integrity of the experiments.

4.2 Evaluation Method

We used the Precision Method [6], Recall Method [6] and F-Measure Method [19] in our work to evaluate the performance of the extraction:

Overall Evaluation: The overall accuracy of classification of the extracted items from the publication data is represented by this formula:

$$Overall = Recall \times \left(2 - \frac{1}{Precision}\right) \tag{1}$$

Class-Specific Evaluation: Specific class means that we performed a separate evaluation for each extracted item: Author, Title, Journal and Year. Assuming that the actual number of phrases in the Training Corpus is A, and the total number of phrases extracted by the system is B, then the precision, recall and F-Measure are defined as follows:

$$Prcision = \frac{|A \cap B|}{|B|} = \frac{\text{the number of correct extracted phrases of the system}}{\text{the total number of extracted phrases of the system}} \quad (2)$$

$$Rrecall = \frac{|A \cap B|}{|A|} = \frac{\text{the number of correct extracted phrases of the system}}{\text{the actual number of phrases}} \quad (3)$$

$$F\text{ - }Measure = \frac{2 \times Precsion \times Recall}{Precisionion + Recall} \quad (4)$$

4.3 Experimental Results and Evaluation

We used the method described in this paper to test the extraction performance of publication data. We extracted four items from the 50 data entries of the Testing Corpus. Table 7 contains the evaluation of these system extraction results. The data show that the overall precision of our bibliographic extraction system is 98.46%.

Table 7 The Precision, Recall, F-Measure and Overall Evaluation

Overall	98.46%		
Class Name	Precision	Recall	F-Measure
Author	100%	100%	100%
Title	99.1%	100%	99.55%
Journal	100%	98.73%	99.36%
Year	100%	96%	97.96%
Average	99.775%	98.68%	

Comparing the overall and specific item efficiencies, the experimental results show that the Recall figures for "Journal" and "Year" are lower than the others. The successful Recall rate of "Journal" is 98.73%, due to the fact that journal names are sometimes recognized as a "Title." instead. Moreover, the successful Recall rate of "Year" is 96%, the lowest of all the ratings. This is because the item "Year" in the bibliographic extraction is easily misjudged as a page number and sometimes is even omitted due to ambiguous formatting. Here are some examples of misjudgments:

(1) "Year" misjudged as "Title:"
Shi-Cheng Liu and Shinfeng D. Lin, **2006**. BCH Code-Based Robust Audio Watermarking in the Cepstrum Domain, Journal of Information Science and Engineering, Vol. E89-D, No. 3, pp. 535-543.

(2) Ambiguous format for "Year:"
Chen, Hong, and Yat-wah Wan, [**2005, May**]. Capacity competition of make-to-order firms, Operations Research Letters, Vol. 33, No. 2, pp. 187-194.

To evaluate the overall efficiency of our method, we made a comparison of our results with Seymore's [20], which is based on HMM, and Peng's [16], which is based on CRF theory. The comparison is illustrated in Table 8. The data show that the overall efficiency of our results is 98.46%, which is better than the overall efficiency of Seymore's results, which is 85.1%. Our overall efficiency is also better than the overall efficiency of the CRF-based method proposed by Peng, which is 95.37%. Furthermore, the data show that our method is superior in the extraction of "Author" items, achieving 100% in both Precision and F-Measure.

To assess the performance of our system, we implemented experiments comparing the number of training data processed, precision and time of training. The results are shown in Table 9. We can see that the precision of the extraction grows correspondingly with the number of training data processed. When the number of training data is more than 90, the precision reaches the highest value.

5 Conclusions and Future Work

This paper presents a bibliographic extraction system, which is based on CRF theory. Using the CRF machine learning technique, we can extract the items of "Author," "Title," "Journal," and "Year" from bibliographic references with high accuracy and produce classified statistics of articles based on a specific purpose.

Table 8 Comparison of Our Results with Seymore's [20] and Peng's [16]

	Seymore [20]		Peng [16]		Our Results	
Overall	85.1%		95.37%		98.46%	
	Pre.	F-m	Pre.	F-m	Pre.	F-m
Author	96.8%	92.7%	99.9%	99.4%	100%	100%
Title	94.4%	85%	97.7%	93.7%	99.1%	99.55%
Journal	96.6%	67.7%	99.1%	91.3%	100%	99.36%
Year	99.7%	96.9%	99.8%	98.9%	100%	97.96%
Average F-m		85.58%		95.83%		99.22%

Table 9 Comparison of the Number of Training Data, Precision and Time

Number	Precision	Time (Sec.)
10	0.5445	1
20	0.5825	2.58
30	0.6468	4.44
40	0.7133	6.06
50	0.81	7.19
60	0.859	9.31
70	0.828	17.3
80	0.99	20.5
90	0.9969	22.34
100	0.99775	26.36

The results of our research show that:

(1) Different from prior research that used the SVM or HMM machine learning technique, we have successfully extracted English publication data based on the CRF theory in this paper. We used the CRF machine learning technique, combined with POS phrase judgment and word marking for identifying bibliographic items correctly.

(2) Our CRF-based automatic bibliographic extraction system exhibited 98.46% overall efficiency, which is better than Seymore’s HMM-based bibliographic extraction method [20], which achieved 85.1% efficiency. Our results are also better than the overall efficiency of Peng’s [16] CRF-based extraction method, which is 95.37%.

(3) As shown in our bibliographic extraction system, the SEM will compare and filter the important information of each research article according to the criteria selected by the user. Thus, the system can be further developed to construct an “Academic Performance Evaluation System for Teachers.” The system can be used to convert teachers’ publication data into statistical evaluation data without the troublesome, time-consuming task of manually counting the number of published articles of every teacher. The system can also be used to provide a flexible system for organizing and managing a personal publication database based on various items. The mechanism provided by the automatic bibliographic extraction system can also be used on a periodical evaluation database for the faculty of a university.

We believe the following work is worthy of investigation in the future:

(1) This paper is focused on analyzing only the bibliographic components of each article of literature. However, the same method and principle can be applied to analyzing the full content of articles by changing the current programs only slightly.

(2) A powerful and comprehensive bibliographic extraction system based on CRF theory is needed for Chinese (and for other languages as well).

(3) An error adjustment mechanism is needed to verify the input article data. Currently, the automatic bibliographic extraction system is dependent upon the user inputting “feasible” and “recognizable” publication data. Data in the wrong format, input errors, and incorrect punctuation decrease the correctness of our system. To raise accuracy, we can use the error adjustment method to help the model learn and adjust the rules for bibliographic extraction.

(4) It would be worthwhile to connect the automatic bibliographic extraction system with some large indexing systems, such as SCI and SSCI, for impact factors extraction and evaluation.

References

[1] Alireza Mansouri, Lilly Suriani Affendy and Ali Mamat, *A New Fuzzy Support Vector Machine Method for Named Entity Recognition*, *Proc. of ICCSIT 2008*, Singapore, August, 2008, pp.24-28.

[2] Baofeng Guo, Steve Gunn, R. I. Damper and James Nelson, *Customizing Kernel Functions for SVM-Based Hyperspectral Image Classification*, *IEEE Transactions on Image Processing*, Vol.17, No.4, 2008, pp.622-629.

[3] Manabu Torii, Kavishwar Wagholikar and Hongfang Liu, *Using Machine Learning for Concept Extraction on Clinical Documents from Multiple Data Sources*, *Journal of American Medical Information Association*, Vol.18, No.5, 2011, pp.580-587

[4] Gend Lal Prajapati, *Advances in Learning Formal Languages*, *Proceedings of IMECS 2011*, Hong Kong, China, March, 2011, http://www.iaeng.org/publication/IMECS2011/IMECS2011_pp118-126.pdf

[5] Hai Leong Chieu and Hwee Tou Ng, *A Maximun Entropy Approach to Information Extraction from Semi-structured and Free Test*, *Proceedings of 18th National Conference on Artificial Intelligence*, Alberta, Canada, July, 2002, pp.786-791.

[6] Cyril W. Cleverdon, *On the Inverse Relationship of Recall and Precision*, *Journal of Documentation*, Vol.28, No.3, 1972, pp.195-201.

[7] Dayne Freitag and Nicholas Kushmerick, *Boosted Wrapper Induction*, *Proc. of AAAI 2000*, Austin, TX, August, 2000, pp.577-583.

[8] Guangxi Chen, Jian Xu and Xiaolin Xiang, *Neighborhood Preprocessing SVM for Large-Scale*

Data Sets Classification, *Proc. of FSKD 2008*, Shandong, China, October, 2008, pp.245-249.

[9] Jun'Ichi Kazama, Takaki Makino, Yoshihiro Ohta and Jun'Ichi Tsujii, *Tuning Support Vector Machines for Biomedical Named Entity Recognition*, *Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain*, Philadelphia, PA, July, 2002, pp.1-8.

[10] Hei-Chia Wang and Tian-Hsiang Huang, *An Enhanced Case-Based Reasoning Model for Supporting Inference Missing Attribute and Its Feature Weight*, *Journal of Internet Technology*, Vol.13, No.1, 2012, pp.45-56.

[11] John Lafferty, John Lafferty and Fernando Pereira, *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*, *Proc. of ICML 2001*, Williamstown, MA, June, 2001, pp.282-289.

[12] Dalei Wu, Yan Yin and Hui Jiang, *Large-Margin Estimation of Hidden Markov Models with Second-Order Cone Programming for Speech Recognition*, *IEEE Transaction of Audio, Speech, and Language Processing*, Vol.19, No.6, 2011, pp.1652-1664.

[13] Kairong Li, Guixiang Chen and Jilin Cheng, *Research on Hidden Markov Model-Based Text Categorization Process*, *International Journal of Digital Content Technology and its Applications*, Vol.5, No.6, 2011, pp.244-251.

[14] Andrew Mccallum and Wei Li, *Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons*, *Proceedings of CoNLL 2003*, Edmonton, Canada, May, 2003, pp.188-199.

[15] Michael Granitzer, Maya Hristakeva, Robert Knight, Kris Jack and Roman Kern, *A Comparison of Layout Based Bibliographic Metadata Extraction Techniques*, *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, Craiove, Romania, June, 2012, doi:10.1145/2254129.2254154.

[16] Fuchun Peng and Andrew Mccallum, *Accurate Information Extraction from Research Papers Using Conditional Random Fields*, *Proc. of HLT-NAACL 2004*, Boston, MA, May, 2004, pp.329-336.

[17] David Pinto, Andrew McCallum, Xing Wei and W. Bruce Croft, *Table Extraction Using Conditional Random Fields*, *Proceedings of SIGIR 2003*, Toronto, Canada, July, 2003, pp.235-242.

[18] R. L. Rabiner, *A Tutorial on Hidden Markov Model and Selected Applications in Speech Recognition*, *Proceedings of the IEEE*, Vol.77, No.2, 1989, pp.257-286.

[19] Manabu Ohta, Daiki Arauchi, Atsuhiro Takasu and Jun Adachi, *CRF-Based Bibliography Extraction from Reference Strings Focusing on Various Token Granularities*, *Proc. of 2012 10th IAPR International Workshop on Document Analysis Systems (DAS)*, Queensland, Australia, March, 2012, pp.276-281.

[20] Kristie Seymore, Andrew Mccallum and Ronald Rosenfeld, *Learning Hidden Markov Model Structure for Information Extraction*, *Proc. of AAAI 99 Workshop on Machine Learning for Information Extraction*, Orlando, FL, July, 1999, pp.37-42.

[21] Fei Sha and Fernando Pereira, *Shallow Parsing with Conditional Random Fields*, *Proc. of NAACL 2003*, Edmonton, Canada, May, 2003, pp.134-141.

[22] Liam Stewart, Xuming He and Richard S. Zeme, *Learning Flexible Features for Conditional Random Fields*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.30, No.8, 2008, pp.1415-1426.

[23] Richard Sproat and Chilin Shih, *A Statistical Method for Finding Word Boundaries in Chinese Text*, *Computer Processing of Chinese and Oriental Languages*, Vol.4, No.4, 1990, pp.336-351.

[24] Vladimir Naumovich Vapnik, *The Nature of Statistical Learning Theory*, New York, Springer, 1995.

[25] Stefan Wermter, Ellen Riloff and Gabriele Scheler, *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, Springer, Berlin, Germany, 1996.

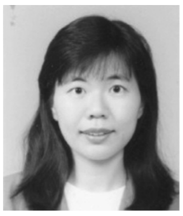
[26] Xiwu Han, Tiejun Zhao and Muyun Yang, *FML-Based SCF Predefinition Learning for Chinese Verbs*, *Proc. of LJCNLP 2004*, Hainan Island, China, March, 2004, pp.664-673.

Biographies

Sheng-Ming Wang is currently an assistant professor in the Graduate Institute of Interactive Media Design, National Taipei University of Technology, Taiwan. He received his BS degree in the Department of Urban Planning, National Cheng Kung University, Taiwan, in 1986, the MS degree in Graduate Institute of Building and Planning from National Taiwan University, Taiwan, in 1992, and the PhD degree in School of Computer Science, University of Leeds, UK, in 1998. Dr. Wang's held multi-disciplinary research interests include information retrieval, data hiding algorithm, geographic information science, interactive media design, serious game design and knowledge management.

Wei-Pang Yang is a professor in the Department of Information Management and the Vice-Principal of National Dong Hwa University, Taiwan. He received the BS degree in Mathematics from National Taiwan Normal University, Taiwan, in 1974, and the MS and PhD degrees from the National Chiao Tung University, Taiwan, in 1979 and 1984 respectively, both in Computer Engineering. His research interests include database theory and application, information retrieval, data mining, digital library, and digital museum. Professor Yang is a senior member of IEEE, and a member of ACM.

Hsin-Ping Chou received the MS degree in Information Management from National Dong Hwa University, Taiwan, in 2007. Now she is working as a database administrator for the Funtown company in Taiwan. Her research interests include database administration, data mining, and knowledge management.



Fu-Mei Chen is currently an associate researcher in the Research and Development Center of National Taipei University of Technology, Taiwan. She received the Economics and Philosophy dual BS degrees from National Taiwan University, Taiwan, in 1992. She received her MBA degree and MEd degree in Information Technology, Multimedia, and Education respectively from the School of Business and Economics Studies, in 1995, and the School of Education, in 1997, both from the University of Leeds, UK. She received the PhD degree in the department of Business Administration, MIS division, National Dong Hwa University, Taiwan, in 2009. Her research interests include data mining, knowledge management, data hiding, business innovation, entrepreneurship and innovation.

Jia-Li Hou is currently an assistant professor in the Department of Information Management of National Dong Hwa University, Taiwan, R.O.C. He received the MBA degree and PhD degrees from the National Central University, Taiwan, R.O.C. Hou's primary research interests include information security, data mining, financial engineering and ERP.

Jyh-Jian Sheu is currently an associate professor in College of Communication, National Chengchi University, Taiwan. He received BS degree in Management Information Systems from National Chengchi University, Taiwan. And he received his MS and PhD degrees in Computer and Information Science from National Chiao Tung University, Taiwan. Sheu's primary research interests include data mining, Internet security, and interconnection networks.