Design and Application of Intelligent Forensics Schema For Mass Unstructured Data

Wenhua Luo
Criminal Investigation Police University of China
Shenyang, China
E-mail: Luowenhua770404@126.com

Kexin Su
Criminal Investigation Police University of China
Shenyang, China
E-mail: 353837331@qq.com

Abstract—**The proportion of unstructured data is much bigger than that of structured data. However, the research carried out in the field of the methods of unstructured data processing and analysis is not as intensive as that of structured data processing and analysis. This paper mainly involves the importance of the research on unstructured data. From the perspective of digital investigation, it further interprets the key techniques used in the processing and analysis of unstructured data, such as naming entity recognition, entity relation extracting, Web data extracting, etc. Combining the self-developed "Intelligent Analyzing System for Mass Case Information" and taking handling the online ball gambling cases in Chinese Mainland as the background in terms of the usage and application, this paper gives detailed explanation to the application of unstructured data processing and analysis in digital investigation in detail.**

Keywords-Intelligent Forensics Software, Mass Unstructured Data; Naming Entity Recognition; Entity Relation Extraction; Online Ball Gambling

I. INTRODUCTION

With the fast development of the technology, the information naturally falls into two categories. One class can be fully expressed and achieved in logical sense by two-dimensional table structure. Thus it can be stored and processed in database, and it is called structured data. The other one cannot be expressed by two-dimensional logical table of database. It includes documents, texts, pictures, web pages, XML, all kinds of reports, images, audio/video information, etc., and it is called unstructured data. In terms of information processing, we will first think of the database. But, in fact, the structured data used by the database to manage and process accounts for 20% of the information totally. It is not satisfying in processing unstructured data with a higher amount[1].

The series of Analyst's Notebook software developed by i2 Corporation in America can help the digital investigator to understand quickly the complicated case and find out the correlation among the mass data unrelated with each other. It can describe and display the correlation among the mass data in the way of graphics in order to find out and disclose the hidden common elements and correlation. However, the disadvantage of Analyst's Notebook series is that it processes structured data of database (the given data should be imported to database according to some kinds of rules). It considers few of unstructured data (all kinds of reports, digital documents, all kinds of elements on Website, pictures, scanned images, large amount of multimedia audio, video information and so on) which accounts for 80% of the total information[2]. Moreover, it cannot identify well the usefulness of the given information.

Because of the above situation, research on handling and processing unstructured data has become one of the important developing trends of computer forensics. Section 2 introduces the relevant technique difficulties and such corresponding solutions of entity relation extracting, Section 3 illustrates "Intelligent Analyzing System for Mass Case Information" developed by both Criminal Investigation Police University of China and Dalian University of Technology to introduce the general framework of the System and the corresponding function module. According to the increasingly rampant online ball gambling crimes in Chinese Mainland, it explains the specific application of "Intelligent Analyzing System for Mass Case Information" in practical cases. Section 4 summarizes this paper, pointing out the drawbacks of relevant work and prospects for the future developing trend.

II. ENTITY RELATION EXTRACTING

The existence of digital evidence provides the possibility of cracking down relevant cases for the digital investigators. However, usually the amount of digital data is very large. The workload of checking digital evidence manually is very heavy[3]. We consider using the method of text mining and knowledge finding to get the entity relation information from the mass digital evidence.

A. Technique Difficulties

Entity relation extracting refers to automatically recognizing the redefining relation between two entities included in the natural language text. Entity relation extracting has significant research meaning in the field of data structuring, data retrieval, automatic answering, etc. Our main goal is to apply the information extracting technique in the NLP(Natural Language Processing), especially the entity relation extracting technique, to the computer relevant crime cases. On the basis of having recognized corresponding naming entity in Section 3, the entity interrelation can be further recognized in order to provide reference for the digital investigators. In fact, in the aspect of entity relation extracting, many scholars have made widely and in-depth research and put forward some effective algorithms. Presently, the widely used one is machine learning method with supervision. The entity relation among untaged data can be forecast through trained extracting model. However, the existing method has the following disadvantages in processing the information of computer relevant crime cases:

The machine learning method depends on specific fields. The feature set must be redesigned for every new field. For the computer relevant crimes, many information is underlying and it is hard to mine out specific features.

Even if the feature set is determined, in the new field some machine learning methods with supervision still need a large amount of manual tag; For computer relevant crimes, the crime information is usually large and various. It is hard that the evaluating corpus has a unified format. The application of tag method to crime information extracting still needs improving[4].

Some naming entity may appear several times in a text. The manifestation may differ (such as pronoun, reflexive pronoun, nominal time presentation and others), so the entity relation is often detected repeatedly. The merging operation of the same entity relation examples is necessary. Besides, the texts with different formats have different ways of displaying the information and have different semantic information. Different forms of data needs different entity relation extracting tasks.

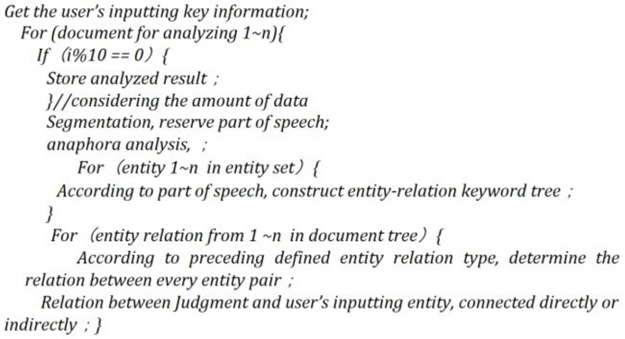
B. Solution

Wenjie L. and others introduces new methods of Chinese naming entity relation extracting in the paper of A Novel Feature-based Approach to Chinese Entity Relation Extraction. What is different from the methods of preceding researchers is that this paper defines nine kinds of different entity relations through nine position relations among entities. They are nested relation, nested-nested relation, coincident relation, adjacent relation, nested-adjacent relation[5], nested-nested-adjacent relation, separated relation, nested-separated relation and nested-nested-separated relation. It is no doubt that putting forward these nine relations contributes much to the research on entity relation extracting. Based on this, we consider practical characteristics of digital investigation work to define eight kinds of relations. The entities are divided into relations of natural persons and relations of business entities[6]. The relation model first defines the relation categories and the situations as much as possible that may be included in these categories. Before defining relation model, we first analyze the corpus extracted from a large amount of true cases and define several relation types that are most likely to encounter (shown in Table 1). Because the definition of entity and relation types should be generalized rather than simplified, we should define them for the cases as much as possible rather than limited to certain types.

TABLE I. TYPE RELATION DEFINITION AND DESCRIPTION

Relation	Description	Representation	Example
R11	A natural person has some qualities.	Per-with-Prop	Go to find Per1, he will be dressed in black jacket.
R12	A natural person knows another natural person.	Per-knows-Per	Per2 and Per3 have meal together.
R13	A natural person is employed in a corporation.	Per-employed-Corp	Per4 of Per4 Corporation
R14	Natural person-contact method	Per-Comm	You may call Tel1 to find Per6.

After defining relation types, we use the following pseudocode to describe entity relation extracting procedure:



```
Get the user's inputting key information;
  For (document for analyzing 1~n){
    If （i%10 == 0）{
      Store analyzed result ；
      }//considering the amount of data
      Segmentation, reserve part of speech;
      anaphora analysis, ；
          For （entity 1~n  in entity set）{
        According to part of speech, construct entity-relation keyword tree ；
      }
       For （entity relation from 1 ~n  in document tree）{
          According to preceding defined entity relation type, determine the
   relation between every entity pair ；
     Relation between Judgment and user's inputting entity, connected directly or
indirectly ；}
```

The construction of a tree is that any two entities are the left son and the right son of a relation keyword. If three or more entities have some relations, the first two entities constructs subtree first. Then the third entity and father node of this subtree construct subtree by general relation keywords. In the same way, the final constructed one is entity relation subtree corresponding to corresponded corpus. The relation keyword is the main basis of extracting entity relation. Different keywords mean different relations among entities. Up till now, conjunction and verb are the relation keywords, such as “and”, “call to”, “find”,etc[7].

III. INTELLIGENT ANALYSIS SYSTEM FOR MASS CASE INFORMATION

A. General Structure

Combining the specific solution methods introduced in Section 2, we developed the “Intelligent Analysis System for Mass Case Information” which focuses on unstructured data and is based on Chinese. This System can conduct division and POS tagging on unstructured case information when preprocessing[8]. According to preset rule library and combining collected context, it can achieve naming entity recognition through the relevant information of user configuration. For the semi-struced turfile, such as the HTML webpage file, achie the structure of case information based on webpveage preprocessing s for solvinand through such operations as HTML analysis and Table data extractor. The above processed result is sent to entity relation extractor module for processing. The entity relation tree is contructed by predefined relation type. Under restricted situation, complete the functions of statistics and calculation, and give feedbacks to the digital investigators in graphic form. (The overall structure of the System is as shown in Figure 1.)

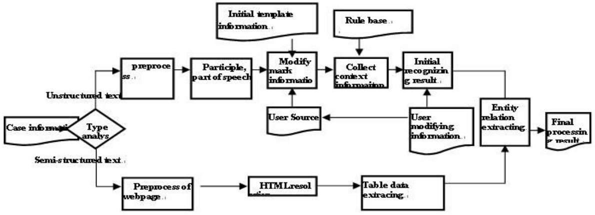


Figure 1. General Structure of Intelligent Analysis System for Mass Case Information

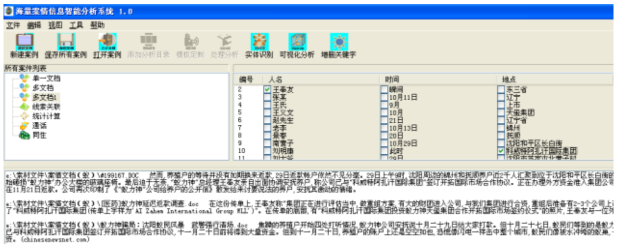
B. Function Instruction

"Intelligent Analysis System for Mass Case Information" can preprocess data item of the mass information source. It can conduct intelligent information analysis through the four following function modules: "Data Statistics and Analysis", "Corpus Recognition Analysis", "Clue Expansion Relation" and "Public Security Basic Data Processing and Analyzing". (The first three modules focus on unstructured data. The last module focuses on structured data.) It can visually output digital evidence or case clues. (The System interface is as shown in Figure 2.) The specific functions of all the modules is as the following:

Through analyzing webpage and fragmented data and after preprocessing, analyzing and data extracting, store extracted data to database. Based on this, it provides visualized "crime feature relation analysis" and conducts statistics and query on case important information. It can serve as the auxiliary tool for relevant case analyzing and processing.

Corpus Recognition and Analysis Module adopt text mining technique to automatically recognize the entity with special meaning from a huge amount of information or text information library, such as person's names, place's names, time, quantity, etc. It can cluster the text according to selected keywords, then form abstract. Through the mrn from information source for the investigators, thus the besatioabove processing, it directly provides important infoetting of redundant information can be avoided and the efficiency of solving a case can be improved.

Clue Expanding Relation Module can emphatically tag the phone number, E-mail, virtual identification number. It adopts data relation technique to sum up the relations, law of different places, all targets and behavior intension. It selects text clue analyzed manually to be keywords. This module can list all relevant person's names, time and acts in the form of graphics to display the case intuitively and restore the original procedure. Thus it provides powerful clueg the case.



编号	人名	时间	地点
2	☑ 王奉友	☐ 瞬间	☐ 东三省
3	☐ 张某	☐ 10月11日	☐ 辽宁
4	☐ 王氏	☐ 9月	☐ 上市
5	☐ 王义文	☐ 10月	☐ 天蛋集团
6	☐ 赵先生	☐ 21日	☐ 辽宁省
7	☐ 老李	☐ 10月13日	☐ 锦州
8	☐ [illegible]	☐ 20日	☐ 抚顺
9	☐ [illegible]	☐ 10月29日	☐ 沈阳和平区长白街
10	☐ 刘明康	☐ 此时	☑ 科威特阿扎汗国际集团

e:\素材文件\案情文档(蚁)\#199167.DOC 然而，养殖户的等待并没有如期换来返款。29日返款帐户依然不见分毫。29日上午9时，沈阳周边的锦州和抚顺养户近2千人汇聚到位于沈阳和平区长白街的蚁力神办公大楼的[illegible]。最后迫于无奈，"蚁力神"总经理王奉友亲自出面协调安抚养户，称公司已与"科威特阿扎汗国际集团"签订开拓国际市场合作协议。正在办理外方资金注入集团公司 在11月21日返款。公司再次印制了《"蚁力神"公司给养户的公开信》散发给来讨要说法的养户，安抚其激动的情绪。

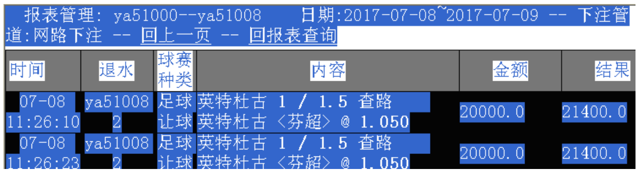
e:\素材文件\案情文档(蚁)\[医药]蚁力神延迟返款调查.doc 在这份传单上，王奉友称"集团正在进行评估当中，做重组方案，有大的财团进入公司，与我们集团进行合资，重组后准备有2-3个公司上 了"科威特阿扎汗国际集团(传单上字样为'Al Zahen International Group KLL')"。在传单的底部，有"科威特阿扎汗国际集团投资蚁力神天蛋集团合作开拓国际市场签约仪式"的照片，王奉友与一位外

e:\素材文件\案情文档(蚁)\蚁力神骗局：沈阳蚁民风暴 武警强行清场.doc 焦躁的养殖户开始四处打听情况，蚁力神公司安抚说十月二十九日给大家打款。但十月二十七日，蚁民们等到的是蚁力 已与科威特阿扎汗国际集团签订开拓国际市场合作协议，十一月二十日前将得到大量资金。但到十一月二十日，养殖户的账户上还是空空如也，恐慌像闪电一样击中整个城市，蚁民们像被水冲垮的蚁巢， 巢。(chinesenewsnet.com)

Figure 2. Interface of Intelligent Analyzing System for Mass Case Information

C. Case Analysis

This section takes a real online ball gambling case as an example to explain the specific application of "Intelligent Analyzing System for Mass Case Information" in the practical work. Because of the national conditions, online ball gambling is illegal in China. From October, 2013 to July, 2017, in the special action of striking online ball gambling, cyber security guarding department and society security worked together to solve the "XXX Special Case" in which the criminals appropriated social security fund of nearly 100,000,000 for online ball gambling. Over ten suspects were arrested and over ten desk-top or laptop computers used by the suspects for online ball gambling were seized. On these computers' harddisks, we found out a huge amount of HTML webpages and webpage fragmentations in GB2312 and BIG5 code relevant to online ball gambling. What is displayed in Figure 3 is one webpage screenshot. The content is the webpage screenshot of statistical statement that member of ya51008 sent to the banker. From the figure, what is displayed is that from 2017-07-08 to 2017-07-09 the member of ya510008 bet twice and the total bet amount is 40000 yuan.



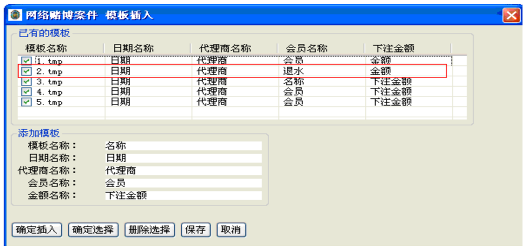
报表管理: ya51000--ya51008 日期:2017-07-08~2017-07-09 -- 下注管道:网路下注 -- 回上一页 -- 回报表查询

时间	退水	球赛种类	内容	金额	结果
07-08 11:26:10	ya51008 2	足球 让球	英特杜古 1 / 1.5 查路 英特杜古 <芬超> @ 1.050	20000.0	21400.0
07-08 11:26:23	ya51008 2	足球 让球	英特杜古 1 / 1.5 查路 英特杜古 <芬超> @ 1.050	20000.0	21400.0

Figure 3. A certain kind of Soccer gambling webpage

Although the most important and direct evidence of the online ball gambling is the webpage which the suspect used for ball gambling bet is found out, it is a sticky problem to calculate such information as the members, agents, bet amounts, data, time of the webpage, etc. Because the quantity of the storage media relevant to computer crime case is large, the capacity is large. The capacity of individual storage media reaches one hundred GB or even several hundred GB. The quantity of information files relevant to the case and searched from these media is a huge amount. It will cost a huge amount of time and energy to analyze and calculate these files manually. Under such circumstances, we used to use the script language of EnScript provided by Encase. The comparatively complicated intelligent processing operations can be achieved through the EnScript programming to save human and material resources. But the script programming still has the following disadvantages: 1) It is very difficult to program. So it is usually suitable for very professional people. 2) It is not universal. Different webpages need different scripts. 3) The result is not intuitive. It is hard for the users to clearly grasp the data.

The data calculating and analyzing module of Intelligent Analyzing System for Mass Case Information can solve such kind of problems better. Generally speaking, the corresponding column names of the information of different types of ball gambling webpages, such as agents, members, bet amount, are different. For example, the corresponding column name of bet amount information may be “bet amount”, “amount” or “stake amount”. Besides, because of such reasons as that some information is covered, for some extracted and recovered webpage fragmentation, individual column name will be misplaced. For example, the corresponding column name of member information may be “agent”. After the digital investigator analyzes the main types of relevant ball gambling webpages, he can make use of the functions provided by the “template customizing” of the data calculating and analyzing module to set the webpage analysis results to be the new template. The setting contents include such corresponding specific structure names of the information as agent, member account, bet amount on the case-relevant webpages so that the System can conduct the entity recognition and relation extracting according to the template. The customized template can be inserted into the existing template library, as is shown in Figure 4.



网络赌博案件 模板插入

已有的模板

模板名称	日期名称	代理商名称	会员名称	下注金额
☑ 1. tmp	日期	代理商	会员	金额
☑ 2. tmp	日期	代理商	退水	金额
☑ 3. tmp	日期	代理商	名称	下注金额
☑ 4. tmp	日期	代理商	会员	下注金额
☑ 5. tmp	日期	代理商	会员	下注金额

添加模板

模板名称：	名称
日期名称：	日期
代理商名称：	代理商
会员名称：	会员
金额名称：	下注金额

Figure 4. Window Content of “Template Customising”

According to the template customized by the digital investigator, the System can process and analyze the imported information file. In view of ball gambling cases, the System can display the relation between the agent account and the member account in form of graphics (as is shown in Figure 5). According to the intensive relation area of the relation view, the digital investigator can have a more directive recognition on the agenting situation of the agents and the participation of the members and make sure of the case breakthrough according to the intensive degree. The System also provides the functions of query and statistics for the following four types of information: agent, member, date and bet amount. The so called query is that the System displays the bet situation of specified agent or member in special time according to the setting modified by the digital investigator (Figure 6 shows the bet list detail of member of ya51000.). The so called statistics is that the System calculates the total bet amount of the specified agent or member in special time. Through the statistic number, the case can be grasped more accurately and quantized.

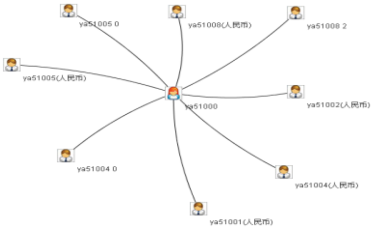
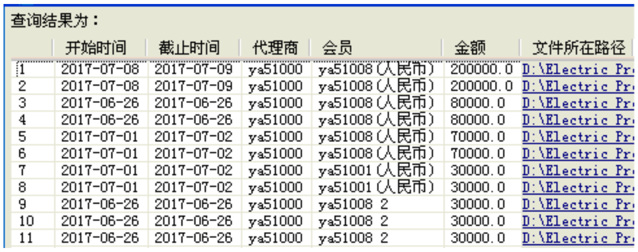


Figure 5. Relation View between Agent and Member



查询结果为：

	开始时间	截止时间	代理商	会员	金额	文件所在路径
1	2017-07-08	2017-07-09	ya51000	ya51008 (人民币)	200000.0	D:\Electric Pr
2	2017-07-08	2017-07-09	ya51000	ya51008 (人民币)	200000.0	D:\Electric Pr
3	2017-06-26	2017-06-26	ya51000	ya51008 (人民币)	80000.0	D:\Electric Pr
4	2017-06-26	2017-06-26	ya51000	ya51008 (人民币)	80000.0	D:\Electric Pr
5	2017-07-01	2017-07-02	ya51000	ya51008 (人民币)	70000.0	D:\Electric Pr
6	2017-07-01	2017-07-02	ya51000	ya51008 (人民币)	70000.0	D:\Electric Pr
7	2017-07-01	2017-07-02	ya51000	ya51001 (人民币)	30000.0	D:\Electric Pr
8	2017-07-01	2017-07-02	ya51000	ya51001 (人民币)	30000.0	D:\Electric Pr
9	2017-06-26	2017-06-26	ya51000	ya51008 2	30000.0	D:\Electric Pr
10	2017-06-26	2017-06-26	ya51000	ya51008 2	30000.0	D:\Electric Pr
11	2017-06-26	2017-06-26	ya51000	ya51008 2	30000.0	D:\Electric Pr

Figure 6. Bet List of Member of ya51000

IV. Conclusions and Future Work

The proportion of unstructured data among the total information is much bigger than the proportion of structured data. The handling and processing of unstructured data is much more difficult than that of the structured data. This paper combines the self-developed “Mass Case Information Analyzing System” to emphatically illustrate the solutions of technique difficulties, such as the naming entity recognition, entity relation extracting, Web information extracting. And it takes cyber ball gambling case handling as the background to focus on the specific application of unstructured data handling and processing in the digital investigation of practical work. So it is a useful exploration and attempt in the field of computer forensics[9].

Before the Intelligent Analyzing System for Mass Case Information analyzes the specific case, it demands the correspondent entity information according to the digital investigator’s understanding about the case. That is to provide the user’s understanding or analyzed case information for the System as the heuristic resource. The System analyzes cases according to this main thread. At the same time, the System uses the user’s input history information stored by the user’s auxiliary resource library. As the investigator who uses Intelligent Analyzing System for Mass Case Information usually has high professional quality, it is of great significance to store their input history information. With the increasing use of the Intelligence Analyzing System for Mass Case Information, the user’s auxiliary resource library will become larger and the recognition efficiency will become higher. Presently, the System is still not perfect, such as that the rule library is not perfect. The extraction of some entity information recognition and relation is not that accurate. Such situations as the same name, alias and the wrong word of the entity can not be recognized. So functions in the aspect of naming entity recognition is still to be perfected, such as the entity disambiguation, cluster analysis, and automatic abstract. On the basis of further enriching rule library resource, the Intelligence Analyzing System for Mass Case Information can provide better service for the investigators to conduct the digital investigation. In the aspect of entity relation extracting, the relevant techniques, such as social network analysis and community structure mining, will be applied to the person's entity relation mining of investigative data. The method of network module can be used to analyze the potential information of crime network in detail. How to show table structure better and make use of application ontology portrait field better will be the focus in the next phase of work of Web information extracting.

ACKNOWLEDGMENT

Funding Support: Key Project of Technical Research Plan of Ministry of Public Security, 2017JSYJA10.

REFERENCES

[1] Sudha Morwal, and Nusrat Jahan. Named Entity Recognition Using Hidden Markov Model(HMM): An Experimental Result on Hindi, Urdu and Marathi Languages. In:International Journal of Advanced Research in Computer Science and Software Engineering,Vol.3, 2013, pp.671-675.

[2] Jana Straková, Milan Straka, and Jan Hajjc. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In:Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2014, pp.13-15.

[3] Yu Hongkui, Zhang Huaping, and Liu Qun. Chinese Naming Entity Recognition based on Layered Hidden Markov Models. Journal on Communications,Vol.27, 2006, pp.53-60.

[4] Chenliang Li, Aixin Sun, and Jianshu Weng. Tweet Segmentation and Its Application to Named Entity Recognition. IEEE Transactions on Knowledge and Data Engineering,Vol.27, 2014, pp.558 - 570.

[5] Min Yang, and Kam-Pui Chow. An Information Extraction Framework For Digital Forensic Investigations. IFIPAICT:Advances in Digital Forensics XI, Vol.462, 2015, pp.61-76.

[6] Zhang Mingde, Bi Maning, Wang Shun. Research on Application Security Formal Description. Network Security, Vol.10, 2016, pp.47-53.

[7] Yang XinYu, Wang Jian. Study on the micro cloud Network Forensics. Network Security, Vol.3, 2015, pp.69-73.

[8] Zhao Jun. Naming Entity Recognition, Disambiguation and Crosslanguage Relation. Chinese Information Journal, Vol.3, No.23, 2009, pp.76-84.