Computer and Information Science

www.ccsenet.org/journal.html

Vol. 2, No. 4

November 2009

Enhancement of Hierarchy Cluster-Tree Routing for Wireless Sensor Network

Xuxing Ding

The College of Physics and Electronic Information, Anhui Normal University, Wuhu 241000, China

Tel: 86-553-388-3560 E-mail: dxx200@163.com

Fangfang Xie

The College of Physics and Electronic Information, Anhui Normal University, Wuhu 241000, China

Tel: 86-553-388-3560 E-mail: fangtinglei@yahoo.com.cn

Qing Wu

The College of Physics and Electronic Information, Anhui Normal University, Wuhu 241000, China

Tel: 86-553-388-3560 E-mail: qq1985918@126.com

The work was supported by University Science Foundation of Anhui Province(No.KJ2009B035) and Key Technology Program of Wuhu (No.2008320)

Abstract

Many protocols such as clustering are proposed to minimize and balance energy consumption of the network because WSN (wireless sensor network) is energy-limited. In clustering protocols, CHs (cluster head) consume much more energy than its CMs (cluster member) which leads to the faster death of CHs. Many traditional protocols are designed to solve the problem, but they have some drawbacks respectively. In this paper, EHCT (enhancement of hierarchy cluster-tree routing for WSN) is proposed to further balance the energy consumption. The simulation results show that the performance of EHCT has an improvement of 41% over LEACH and 14% over UDACH in the area of 500m*500m, 28% over LEACH and 18% over UDACH in the area of 1000m*1000m.

Keywords: WSN, Cluster-tree enhancement, EHCT

1. Introduction

Fast development of technologies such as low-power wireless communication and inbuilt computing enables that many low-price sensors can be organized to be a WSN (wireless sensor network). The sensors are distributed in the open and powered by batteries. They need to work lots of months and can not be powered again. By this token, the difference of WSN from traditional networks is energy-limited. To solve this problem, many clustering protocols have been proposed in recent years.

LEACH (Heinzelman W., 2000, pp.1-10) is a representative clustering protocol where CHs are selected randomly and rotated in each round. The energy consumption is reduced because of the decreasing of the number of nodes which communicate with BS directly.

LEACH-C (Heinzelman W., 2002, 4, pp.660-670) is an improvement of LEACH. In this protocol, CHs are not selected randomly but according to the residual energy. The node whose residual energy is more than the average energy of all nodes may be selected as a CH, so CHs can not die untimely. However, the deficiency of LEACH and LEACH-C is that the distribution of CHs is uneven and the data is transmitted by single hop.

PEGASIS (Lindsey S., 2002, 3, pp.1-6) makes all nodes form a chain and specifies one node to communicate with BS directly. Each node only communicates with its neighbors and sends data to BS in turn, but it requires the location of all nodes.

UDACH (Chen Jing etc., 2007, pp.628-632), a uniformly distributed adaptive clustering hierarchy routing protocol, enables that CHs are uniformly distributed, which balances the energy consumption and prolongs the lifetime of the network compared with LEACH and LEACH-C. But CHs in UDACH communicate with each other directly, when the

distance of two CHs is long, the sending CH consumes much energy.

In this paper, EHCT is designed to further balance energy consumption and prolong lifetime of the network. It selects CHs based on a Master/Slave method. When a CH needs to send data, it selects a RN which is a member of the previous cluster to forward data to the next CH. The total distance of RN to previous CH and RN to next CH is minimum and less than the distance of previous CH to next CH.

2. Wireless network model

The sensor nodes are randomly distributed in a square area and monitor the environment unceasingly. We assume that all nodes are energy-limited. The location of each node is unknown and fixed. The BS is stationary and located far away from the monitored area. Each link is symmetric and the approximate distance of two nodes can be evaluated according to the received signal intensity.

We use the energy model (HeinzelmanW., 2000, pp.83-86) to analyze the radio energy consumption (Figure.1). Two channel models are used in the energy model. One is free space model, the other is multi-path fading model. The distance between transmitter and receiver determines which model is used.

d is transmitting distance of L-bit packet, multi-path fading model is selected when d is greater than d_0, and the energy spent for the radio is

$$E_{TX}(l,d) = l * E_{elec} + l * \varepsilon_{amp} d^4 \tag{1}$$

Free space model is selected when d is less than d_0, and the energy consumption is

$$E_{TX}(l,d) = l * E_{elec} + l * \varepsilon_{fs} d^2 \tag{2}$$

E_{elec} in equation (1) and (2) represents the electronic energy, ε_{fs} and ε_{amp} are transmitter amplifier, d_0 is a constant.

The energy for receiving this packet is

$$E_{RX}(l) = l * E_{elec} \tag{3}$$

The energy for fusing n packets with L-bit is

$$E_{DA}(n,l) = n * l * E_f \tag{4}$$

E_f in equation (4) is the energy consumed by the node to fuse one bit.

3. EHCT design

An enhancement of hierarchy cluster-tree routing EHCT is proposed in this paper. It is an energy efficient clustering protocol which is divided into several rounds. Each round is composed of three phases: cluster formation, cluster-tree construction, data transmission. CHs are selected in cluster formation phase and aggregate the data collected from its CMs in data transmission phase. If the cluster-tree is constructed, CHs forward the aggregated data to BS.

At the beginning of the network construction, BS broadcasts a HELLO packet to all the nodes at a certain power level. Each node can compute the approximate distance to BS according to the received signal intensity.

3.1 Cluster Formation

The cluster formation is based on a Master/Slave method. There is a MCH (master CH) and two SCHs (slave CHs) in each cluster. MCHs are not selected randomly but according to the residual energy. Each node generates a random number and compares it with a specified threshold. The node whose random number is greater than the threshold becomes a candidate MCH. Only the candidate MCH has the right to be a MCH. The candidate MCH which has the max residual energy is the final MCH. Each MCH then selects two SCHs which have the maximum residual energy among its neighbors. The two SCHs then join this cluster. Other neighbors join clusters according to the principle of proximity.

After each cluster is constructed, the MCH sends the TDMA packet to each CM to assign the slot time. The two SCHs have the last two slots. CMs send collected data to the closest one among the MCH and two SCHs.

3.2 Cluster-tree Construction

A cluster-tree is built to link MCHs with BS in this phase. Each MCH broadcasts a WEIGHT packet including its own ID and the square of d(CH, BS) in a certain power level. The node receiving the WEIGHT packet computes the approximate distance to the sending MCH based on the received signal intensity.

We introduce a distance function f(d) to select the next MCH to BS. If the MCH_x receives a WEIGHT packet from MCH_y and $d^2(MCH_x,BS)$ is great than $d^2(MCH_y,BS)$, it computes $f^{xy}(d)$ and selects the MCH which has the minimum f(d) to be the next hop to BS.

$$f^{xy}(d) = \left| d^2(MCH_x, BS) - d^2(MCH_y, BS) \right|, (x, y = 0, 1, \ldots N) \tag{5}$$

N is the number of nodes. When a MCH can not communicate with any other MCH in its transmission range, it sends the data to BS directly.

The MCH which has selected the next MCH further selects a RN to forward the aggregated data. After a certain time, each CM sends a DIST packet to its MCH in its assigned slot time. The DIST packet includes the square of the distance of the CM to its near MCHs.

We assume that MCH_m selects MCH_n to be the next hop. If a CM_i whose MCH is m receives a WEIGHT packet from a MCH_n, it computes the approximate distance $d(CM_i->MCH_m, MCH_n)$. If a MCH_m receives a WEIGHT packet from a MCH_n, it computes the approximate distance $d(MCH_m, MCH_n)$.

We give the following formula:

$$d^2_{comp}(CM_i) = d^2(CM_i \to MCH_m, MCH_n) + d^2(CM_i \to MCH_m, MCH_m) \tag{6}$$

The MCH_m compares d^2_{comp} of all its CMs and selects the CM which has the minimum d^2_{comp} to be the candidate RN. Assuming that CM_j is selected and $d^2_{comp}(CM_j)$ is less than $d^2_{comp}(MCH_m,MCH_n)$, which is

$$d^2_{comp}(CM_j) < d^2(MCH_m, MCH_n) \tag{7}$$

CM_j becomes the RN to forward the data from MCH_m to MCH_n. Otherwise MCH_m forwards the aggregated data to its next MCH_n directly.

Our method of building cluster-tree can decrease the energy consumption of the MCHs and balance the energy consumption of the whole network on the ground that the energy consumption is relative to square distance in free space or four-square distance in multi-path fading space. The flowchart is shown in figure.2.

3.3 Data transmission

Each node collects the data unceasingly and sends it to its MCH or two SCHs in its transmission slot time. The two SCHs send the data aggregated by them to the MCH in the last two slots. The MCH aggregates all the received packets to one single packet and then sends it to its RN which forwards the aggregated packet to the next MCH. If there is not an eligible RN, the MCH sends the packet to next MCH directly.

4. Evaluated performances

To evaluate the performance of EHCT and compare it with LEACH and UDACH, the protocols are simulated in two environments 500m*500m and 1000m*1000m, the simulation tool is OMNET. There are 100 nodes and the BS is located far away from the monitored area. The simulation parameters are listed in table 1.

We analyze the three protocols from two aspects: total energy consumption and death time of half nodes. Figure.3 and figure.4 show the total energy consumption over time in 500m*500m and 1000m*1000m respectively. The simulation results show that EHCT consumes the least energy and UDACH takes the second place, while LEACH consumes maximum energy. CHs in LEACH communicate with BS and the distribution of CHs is uneven. CHs in UDACH communicate with each other directly. The distance between the two CHs is longer, and the energy consumption of the sending CH will be more. In EHCT, we select a CM in the previous cluster to be a RN which forwards the aggregated data from the previous CH to the next CH. It decreases the energy consumption of CHs, thus it reduces the energy consumption of the whole network.

Figure.5 and figure.6 illustrate the number of dead nodes over time in 500m*500m and 1000m*1000m respectively. The simulation results show that the fewer the dead nodes are, the better the performance of the protocol is. As shown in the two figures, LEACH has the worst performance while EHCT is best.

We usually evaluate the performance of a protocol based on the death time of half number of nodes which represents the lifetime of network. As shown in table 2, the longer the death time of half number of nodes is, the better the protocol is.

The performance of our protocol compared with UDACH and LEACH is shown in table 3. The performance of EHCT has an improvement of 41% over LEACH and 14% over UDACH in the area of 500m*500m, and 28% over LEACH and 18% over UDACH in the area of 1000m*1000m.

The simulation results show that EHCT balances the energy consumption of the nodes and prolongs the lifetime of network. It performs best of the three protocols.

5. Conclusion

To balance the energy consumption and prolong the lifetime of the network, EHCT is proposed. It is composed of three phases: cluster formation, cluster-tree construction and data transmission. After clusters are built, a previous MCH further away from BS may forward the aggregated data to a RN, and then the RN forwards the data to the next MCH nearer to BS. The RN is a CM whose MCH is the previous MCH. The simulation results show that EHCT has the best performance compared with LEACH and UDACH.

References

Chen Jing, & Yu Fengqi. (2007). An Uniformly Distributed Adaptive Clustering Hierarchy Routing Protocol[C]. *Proceedings of the 2007 IEEE International Conference on Integration Technology*, Shenzhen, China, 628-632.

Heinzelman W, Chandrakasan A, & Balakrishnan H. (2000). Energy-Efficient Communication Protocol for Wireless Microsensor Networks. *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, Maui, Hawaii, USA, 8, 1-10.

Heinzelman W. (2000). Application-Specific protocol architectures for wireless networks. Ph. D. Thesis. Boston : *Massachusetts Institute of Technology*.

Heinzelman W., Chandrakasan A., & Balakrishnan H. (2002). An Application-Specific Protocol Architecture for Wireless Microsensor Networks. *IEEE Transaction on Wireless Communications*, 4, 660-670.

Lindsey S., & Raghavendra C. S. (2002). PEGASIS: Power-Efficient Gathering in Sensor Information Systems, *IEEE Aerospace Conference Proceedings*, 3, 1-6.

Younis O., & Fahmy S. (2004). HEED: A Hybrid, Energy-Efficient, Distributed Clustering Approach for Ad hoc Sensor Networks. *IEEE Transactions on Mobile Computing*, 3(4), 366-379.

Table 1. Simulation parameters

Parameter	Value
Initial Energy	200J
Transmitting Radius	150m
Simulation Time	36000s
ε_{fs}	10pJ/bit/m^2
ε_{amp}	0.0013pJ/bit/m^4
E_{elec}	50nJ/bit
E_f	5nJ/bit/signal

Table 2. Death time of the three protocols

Area	Algorithm	Death time
500m*500m	EHCT	31069s
	UDACH	26692s
	LEACH	18198s
1000m*1000m	EHCT	25457s
	UDACH	20960s
	LEACH	18437s

Table 3. Improvement of EHCT

Area	Comparer	Improvement
500m*500m	UDACH	14%
	LEACH	41%
1000m*1000m	UDACH	18%
	LEACH	28%

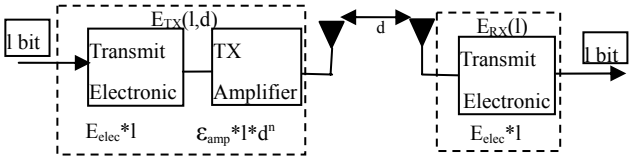

Figure 1. Energy Model

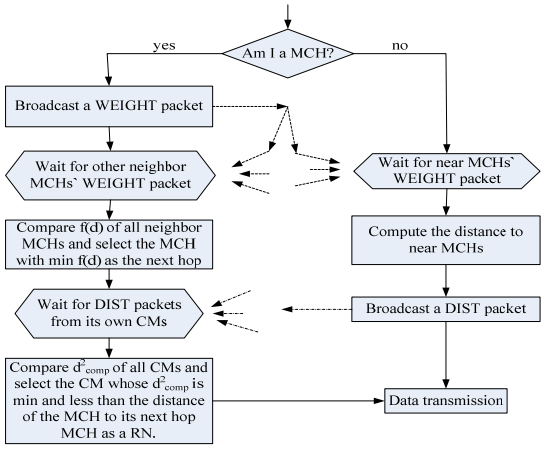

Figure 2. Flowchart of cluster-tree construction

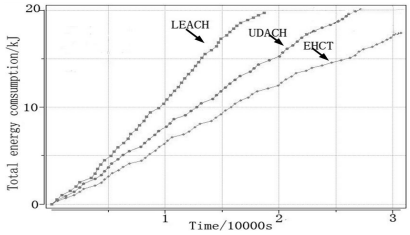

Figure 3.Total energy consumption in 500m*500m

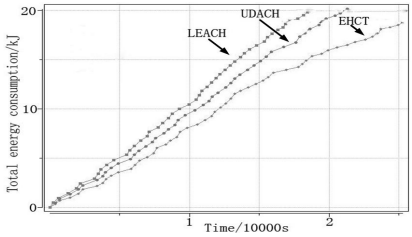

Figure 4. Total energy consumption in 1000m*1000m

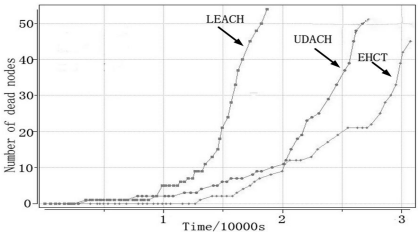

Figure 5. Number of dead nodes in 500m*500m

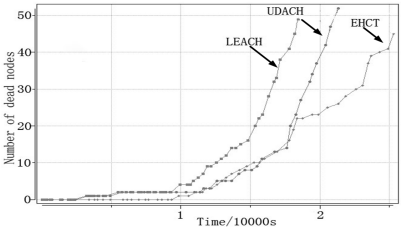

Figure 6. Number of dead nodes in 1000m*1000m

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 2, No. 4

November 2009

Applying Knowledge Management System Architecture in Software Maintenance Environment

Rossey Ginsawat, Rusli Abdullah & Mohd Zali Mohd Nor

Faculty of Computer Science and Information Technology, Universiti Putra Malaysia

43400 UPM Serdang, Selangor, Malaysia

E-mail: rossey_ginsawat@yahoo.com, rusli@fsktm.upm.edu.my

Abstract

Knowledge management (KM) become important for organization to take advantage on the information produced and can be brought to bear on present decision. Software maintenance (SM) is a process that requires lots of knowledge. Maintainers must know what changes should do to the software, where to do those changes and how those changes can affect other modules of the system. Knowledge management system (KMS) can support the processes of knowledge creation, storage or retrieval, transfer and application. KMS in SM could help the organization to make tacit knowledge into explicit and therefore decrease the dependency on employees' cognition. This paper is to apply KMS architecture in SM environment to overcome the problem faced by software maintainers during the software maintenance process.

1. Introduction

KM can be used as a recognition that employees in an organization, as part of their daily activities, knowledge that valuable to the organization. SM is an activity that modifying an existing system to adapt it to new needs, adapt it to never changing environment, or to correct errors in it either preventively or as the result of the actual problem of maintenance project. During SM, maintainers need knowledge of the system they work on, of its application domain, of the organization using it, of past and present software engineering practices, of different programming languages, programming skills. Among the different knowledge needs, maintainers must identify the following:

- Knowledge about the system maintained emerges as a prominent necessity.
- The design decision of the knowledge about software development applied in the process of transforming the knowledge on the application domain into a production of a source code.

KMS could satisfy the needs as to avoid issues in SM. This is because organization can store information and knowledge in KMS. Therefore, even though experts left the organization, the organizations whose own their expertise will remain and maintain the knowledge within the organization. For these reasons, KMS is important so that diverse kinds of knowledge generated from stages of software management process could be stored, analyzed, shared and reused.

Another benefit of KMS is that staff may also inform about the location of the information. Normally, a critical factor for software engineers in maintenance process is that to ensure they will get access to the right knowledge that became the treasure by the organization. Basically the number one barrier to knowledge sharing is ignorance (Szulanki, 1995; Ruiz, *et al,* 2004).

Nowadays, it is an increasing emphasis on the development of information system designed mainly to assist the sharing and integration of knowledge. Most of researcher agreed that KMS architecture should have at least three tiered architecture. They are interface, applications and repositories (Chua, 2004).

When implementing KMS in SM, normally it will involve three profiles of people. These three profiles of people can be clearly differentiated in maintenance process as the maintainer, the administrator, and the user [Collier, 1996].

2. Related Work

A. KM-Mantis

KM-MANTIS is a multi-agent system that was developed to manage knowledge in SM [Palade, 2003]. The architecture used in this system has different types of agent to manage diverse types of information generated during software maintenance process. In this work, agents interchange data and take advantage of the information and experience by other agents.

MAPROK

MAPROK (a multi-agent architecture to process knowledge) uses specialized autonomous agents for specific services and allows agents to interact in order to support the main knowledge processes [Soto, 2006]. The architecture was designed to cater the main processes of knowledge life cycle. In this work the author emphasized that there is generic multi agent architecture to be taken into account when developing KMS such as creating, maintaining, sharing and distributing knowledge.

3. KINOWLEDGE MANAGEMENT SYSTEM ARCHITECTURE IN SOFTWARE MAINTENANCE

The system model consist of three tiered architecture; interface, application and repositories. The system architecture is shown in Fig.1.

A. User Interface Module

User interface module is the front end of this system. The main purpose of this module is to provide the interface to the user for them to communicate with the system. From the interface, the user can key-in their request and then send the request to the application to be processed. Besides that, the interface module also will display the output from the system to the user.

B. Application Module

The system function will reside in application module. The user can find solution and submit their problem and contribution in this module. Some of the tools available in the system are:

i. Submit incident or contribution

ii. Viewing of reward point for each problem solved

iii. Searching for solution history

C. Data access module

Data access module is located between the database and the application module. The main purpose of this module is to connect the application to the database. In this study, the author uses the direct connection to connect the application to the MySQL database.

4. Implementation

KMS in SM is an application that is to be used for creating, storing and reusing knowledge that has been contributed by the maintainers. This work had applied the KMS architecture which was proposed by Chua (2004). The KMS architecture is shown in Fig.2

Explanations for each layer are as follows:

1) Infrastructure Services

This layer comprise of two services that are storage and communication. The storage is to store the knowledge while a communication service is for collaboration among users. But, in this work communication were not covered since it is out of scope.

2) Knowledge Services

This layer comprise of three services. First, knowledge creation which is the creation of knowledge either it is through exploitation, exploration or codification. Second, knowledge sharing is to foster the flow of knowledge among the organization members. Third, knowledge reuse is to capture, package, distribute knowledge and use knowledge.

3) Presentation Services

This service covers the personalization which involves gathering user information and delivering appropriate content. While, visualization is to help users better understand the information and knowledge available by making browsing and navigation easier.

By implementing this architecture, it can compelling the need for KM in organization by a host of social, economic and technological factors including the shortening of product life cycle, fluidity of workforce and the prevalence of work arrangements. Furthermore, when used in tandem with an appropriate KM strategy, technology is a powerful enabler of organizational success. Besides that, through the simplicity of a three-tiered structure, the KMS architecture can help the organization to understand the KM technology.

In this work, users are required to send their request for solution to HelpDesk. Then he will submit the request to Project Manager (PM). PM will categorize the problem whether it is adaptive, corrective, perfective and preventive. Besides that, PM also has to set the problem priority whether it is low, high or normal. After setting both category and priority, PM will submit the task to maintainer. Maintainer will give a solution and submit it back to PM for approval. If the solution is approved by PM, it will be stored in the KMS. Otherwise, PM can always be allowed to resubmit the task or request to maintainers for other resolution. The system flow is shown in Fig. 3.

5. Discussion

Basically, the architecture applied in this system fulfils its objective and successfully completed according to the research scope. The application of KMS architecture in SM has proven that knowledge creation, knowledge sharing and knowledge reuse could be done. Besides that, storage which is also known as knowledge repository can be used to store the knowledge contributed by the user. By having this structure, experience from previous project can always be kept in the KMS and can be referred anywhere and anytime by the organization staff. Additionally, the organization does not have to worry even their staff left the organization because their tacit knowledge will always remain in the organization.

From the survey that we have been conducted through a selected government agencies which is involving in maintaining applications for their businesses as well as having all the criteria given, the respondents were asked on whether the KMS architecture has been applied in the system. From the survey conducted, 57.1% respondent agrees and 42.9% respondent strongly agreed that the system should have collaborative environment. But unfortunately, this feature was not included since it is out of system scope. This information could be used for future work. Additionally, in terms of knowledge services characteristics layer which consist of knowledge creation, knowledge sharing and knowledge reuse, 42.9% respondent agreed and 57.1% respondent strongly agreed that the system had covered this layer. For presentation services, most of the respondent agreed that when visualizing they system they can understand what the system is trying to explain. The percentage that agreed this system had applied this layer is 61.9%. Overall, 57.1% respondent agreed and 19.0% strongly agreed that KMS architecture has been applied in the system.

Fig.4 illustrated the respondent acceptance on KMS architecture.

From Fig.4, it can be viewed that the KMS architecture's layers that are repository, application and interface must be included in implementing KMS.

For infrastructure service that is for storage, 95.2% respondent agreed that the system fulfil this layer. This service is illustrated more in Fig.5.

From the study, it was also found out that even though the KMS architecture model distinctly illustrates various services supported by technology, delineation among services may sometimes be fuzzy. To overcome this weakness, organization is advised to use KMS architecture that has a technology solution merely for its extendibility, comprehensiveness functionalities and technical features. For example, if an organization aims to gain knowledge from users or create knowledge for users, a technology solution that primarily supports the knowledge creation process is preferred to one that supports only the knowledge sharing process.

In this system, users are required to send their request for solution or idea contribution through the knowledge services layer. After all the request is solved, it will be stored in the KMS for future reference. The evaluation done by twenty one respondents (maintainers, administrator, and customers) as mentioned as above, agreed that the developed system fulfil the need of KMS in SM. From the survey that is using the questionnaires as listed the items as below [Item (a) until (d)], we have found that the KMS in SM is very helpful for several purposes:

a) **Reduce time to find expert** which is time consuming. It is also a good tool for sharing pieces of code, patterns, and reusable components with others.

b) **Reward System** is encouraging people to contribute knowledge.

c) **The stored knowledge** can be analyzed and historical data can be keep late it be from the past for future reference.

d) **KMS Architecture** plays an important role to ensure a successful implementation of SM.

6. Conclusion and Future Work

Knowledge is a crucial resource for organization. It allows companies to fulfil their mission and become more competitive. The management of knowledge and how it can be applied to software development and maintenance has received little attention from software engineering research community so far. However, software organization generates a huge amount of knowledge that should be stored and processed. In this way, they would obtain more benefits from it.

The main contribution of this study is the application of KMS architecture in software maintenance environment which considers the processes of a knowledge life cycle such as creating, maintaining, sharing and distributing knowledge. Furthermore, the study also in charge of storing and managing information, expertise and lessons learned which are generated during the software maintenance process. The system facilitates the reuse of good solutions and the sharing of lessons learned. Thereby, the costs in time and effort should decrease. Additionally, in this study, the author has applied the architecture proposed by Chua (2004) which has the interface, application and repository layer.

The project implementation has initiated some ideas in setting up the KMS in SM, as a knowledge repository for the new maintainer or programmer in an organization. The implementation of the KMS in SM may vary from organization to organization. It can be further continuously refined and evaluated with the real user's participation from each cycle in

software development process. It is hope that real user evaluation will provide more valuable input to the refinery of KMS in SM. KMS can be a source of knowledge that can be used by new comer of an organization in finding solution to their problem during SD. So that previous mistake won't be repeated. By referring to KMS in SM this can help the management to reduce cost in hiring expert to solve similar problem faced during previous project.

The main contribution of this project is to apply KMS architecture in SM and therefore build a central repository for SM. The prototype system is a type of knowledge portal (k-portal) only provides an initial demonstration applying the architecture of KMS for SM. Therefore, for future development there is a need to focus on developing an autonomous agent which in charge of giving an appropriate electronic format to experiences obtained so that they can be stored in a knowledge base to aid retrieval. Besides that future work should have an agent which able to collect information for example data, models and experience from different knowledge sources. Furthermore, searcher agent should also be included into a k-portal in order to take charge and produce a recommendations or suggestion with certain goal of helping users to perform their tasks by reusing lessons that is already learnt. Additionally, future development also should include the collaborative environment where expert can refined the available solution that has been already in the system at anytime and anywhere.

References

Chua A. (2004). "Knowledge management system architecture: a bridge between KM consultants and technologies", *International Journal of Information Management* 24, 87-98.

Collier B., DeMarco T., Feary P. (1996). "A defined process for postmortem review", *IEEE Software*. 13 (4), 65-72.

Soto J.P., Vizcaino A., Portill J. o., Piattini M. (2006). "MAPROK: A Multi-Agent Architecture to Develop Knowledge Management Systems", *Proceedings of the Fourth IASTED International Conference Knowledge Sharing and Collaborative Engineering.*

Palade V., Howlett R.J., and Jain L.C. (2003). KES 2003, *LNAI* 2773, pp.415-421, , *Springer-Verlag Berlin Heidelberg.*

Szulanki G. (1994). "Intra-Firm Transfer of Best Practices Project", *American Productivity and Quality Centre, Houston, Texas*, 2-19.

Ruiz, F., Vizcaíno, A., Piattini, M., García, F. (2004). An Ontology for the Management of Software Maintenance Projects. *International Journal of Software Engineering and Knowledge Engineering*, Springer Berlin Heidelberg.

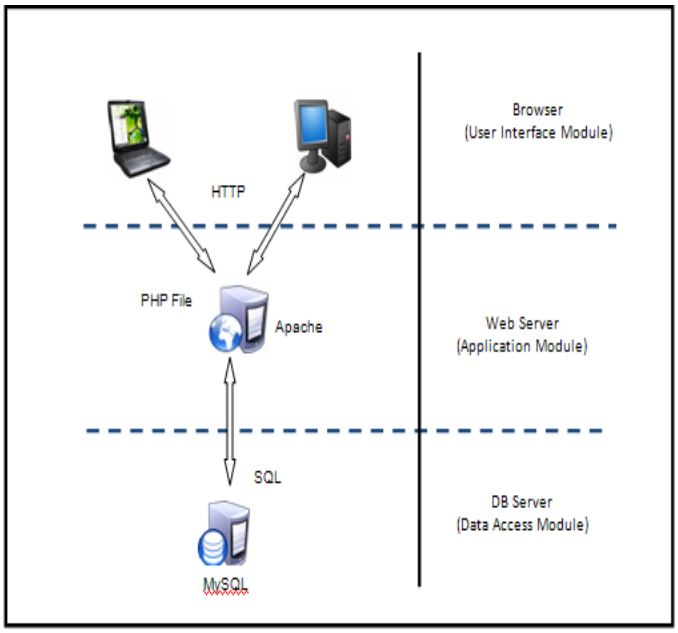

Figure 1. The System Architecture

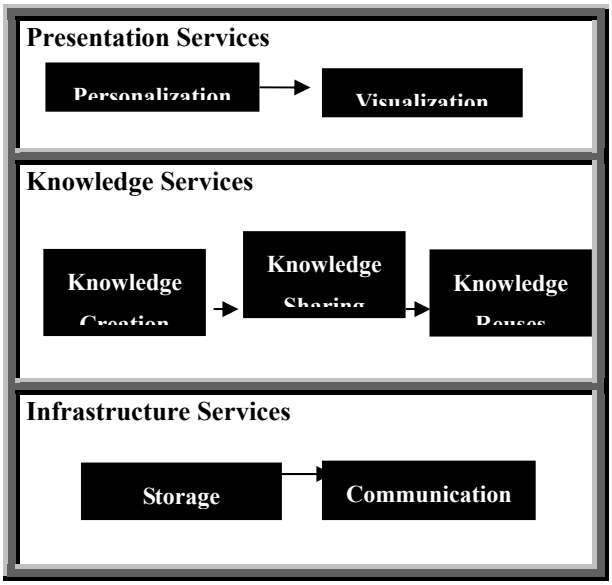

Figure 2. KMS Architecture

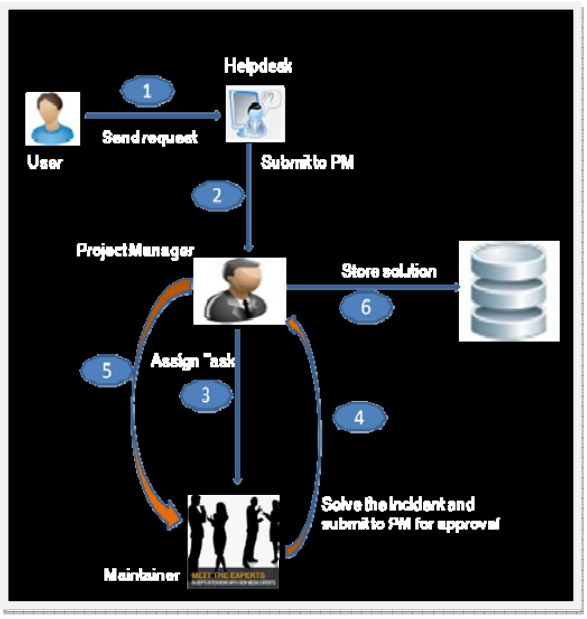

Figure 3. System Flow

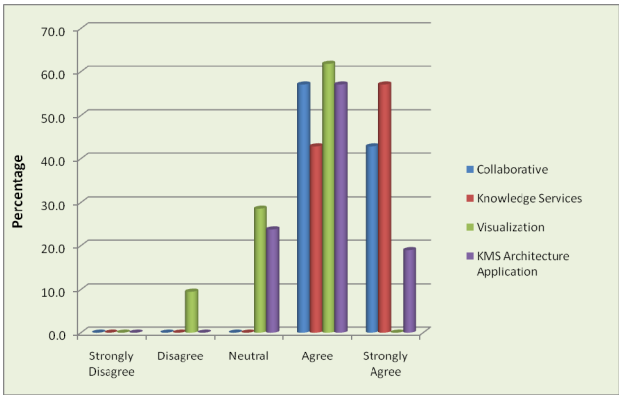

Figure 4. KMS Architecture Acceptance

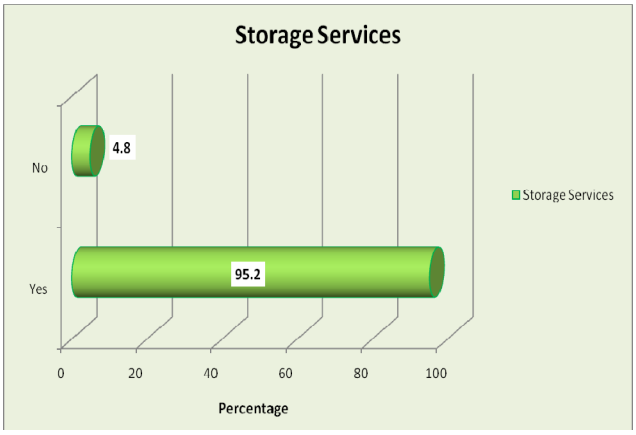

Figure 5. Storage Services

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 2, No. 4

November 2009

Intrusion Detection Method Using Protocol Classification and Rough Set Based Support Vector Machine

Xunyi Ren

College of Computer Science, Nanjing University of Post & Telecommunications

Nanjing 21003, China

E-mail: renxy@njupt.edu.cn

Ruchuan Wang

College of Computer Science, Nanjing University of Post & Telecommunications

Nanjing 21003, China

State Key Lab. for Novel Software Technology, Nanjing University, Nanjing 210093, China

Hejun Zhou

College of Computer Science, Nanjing University of Post & Telecommunications

Nanjing 21003, China

Abstract

In order to improve the efficiency of support vector intrusion detection, we first do protocol Classification for the intrusion data, then refine its characteristic by rough set reduction. By using these procedures, we propose an intrusion detection method using protocol classification and rough set based support vector machine. The method is divided into training and testing processes. In the process of training, we first do protocol classification for the training data, and then do rough set refinement. The refined characteristics are stored as the pre-defined process, and finally the usage of support vector machine for data reduction training, the training model will be stored in accordance with the agreement. In the testing process, the data is classified according to protocol classification and then start the characteristics reduction procedure according to protocol classification. Finally, make a decision using the Support Vector Machines that corresponding to the agreement. The experimental results based on KDDCUP'99 data show that the method is the method is faster and the detection accuracy is comparable compared with the SVM without using protocol classification and using all characteristic.

Keywords: Intrusion detection, Support Vector Machine, Rough set

1. Introduction

Support Vector Machine, refer to Vapnik, 1995, Burges, 1996, P.121-167, is based on structured risk minimization and statistic theory. It overcomes the shortcoming such as difficult to handle of small samples, high dimension, over-matching, local minimization problems etc, that exists in the conventional methods like natural network. Therefore, it is a new high performance learning method, and it has been widely applied in intrusion detection face reorganization, voice processing and so on.

Intrusion detection is essentially a classification problem. It can be viewed as a classification process for test samples of training models. However, the construction of intrusion detection model needs to do learning for thousands of samples; there are tens of characteristics for every sample. Moreover, samples have the property of different structure. If we put the entire characteristic into intrusion detection, SVM will have to solve complex a quadratic programming problem. Therefore, the method is inefficient.

Actually, certain dependent relationship exists in the high dimension characteristics, therefore, how to find this dependence, and then compress the data so as to reduce the dimension, are significant for shorten SVM training time, detection time, and choosing the optimal parameter (Mukkamala, Janoski &.Sung, 2001, P. 1702-1707, Sung, 405-411, Lin & .Cunningham, P.190-198). In (Frohlich, Chapelle & B.Scholkopf, 2003, P.142-148), the genetic algorithm is

adopted to optimize the model and characteristic chosen. In (Roberto, Guofei & Wenke, ICDM'06), n-grams is chosen to choose the host computer character and construct a combined SVM detector. In (Sung, P.405-411), the weighted SVM W is adopted to order and choose the characteristic, and by deleting the low influenced characteristics so as to find the most efficient two kinds of methods.

These kinds of methods have made considerable progress; however, these methods are always distilling the characteristic from all the data. Actually, the intrusion detection usually uses the leak of the protocol, and for every kind of protocol, the intrusion data has different characteristic. For different protocols, if different characteristic is used, the method will more powerful, and hence it will be helpful for improving the learning efficiency of the model.

Rough sets (Pawlak, 1982, P.341-356) is a frequently used method for distilling the characteristics, it is efficient in decreasing the dimension of data. In this paper, we propose to combine the protocol classification and rough sets methods, and so as to produce a intrusion detection method that is based on protocol classification and rough set SVM. By classifying the data based on data protocol, and reduction, we can give the training and detection model. Using the KDDCUP'99 intrusion data, we verify the method.

2. Classify Support Vector Machine (Vapnik, 1995, Burges, 1998, P.121-167)

Suppose $\{(x_1, y_1),(x_2, y_2),\cdots,(x_n, y_n)\}$ is a group of sample data, with $x_i \in \Re^d, y_i \in \{-1,+1\}$. We want to find the optimal partition plane $y = W \cdot X + b$, which is equivalent to solve the convex quadratic Burges programming problem:

$$\begin{aligned} &\min_{w,b,\xi_i} imize && \frac{1}{2} w^T w + C\sum_{i=1}^{N} \xi_i \\ &st && y_i(w^T X_i + b) + \xi_i - 1 \geq 0, \xi_i \geq 0, 1 \leq i \leq N \end{aligned} \tag{1}$$

Where w is the normal vector of hyperplane, b is the deviation, while C a punish function parameter in the case of incomplete integral, and ξ_i is a relaxation parameter in the case of relaxing the constraint conditions. By introducing the Lagrange multiplier:

$$L_p = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} a_i y_i (w^T X_i + b) + \sum_{i=1}^{n} a_i - \sum_{i}^{n} a_i \xi_i \tag{2}$$

Then do partial differential for L_p:

$$\begin{cases} \dfrac{\partial Lp}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{n} a_i x_i y_i \\ \dfrac{\partial Lp}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} a_i y_i = 0 \\ \dfrac{\partial Lp}{\partial \xi_i} = 0 \Rightarrow 0 \leq a_i \leq c \end{cases} \tag{3}$$

In order to obtain a_i, we convert the original problem in to a dual problem, and introduce kernel function $K(\bullet,\bullet)$:

$$\begin{aligned} &\max imize && Q_D = \sum_i a_i - \frac{1}{2}\sum_{i,j} a_i a_j y_i y_j K(X_i, X_j) \\ &st && 0 \leq a_i \leq C, \sum_{i=1}^{n} a_i y_i = 0, 1 \leq i \leq N \end{aligned} \tag{4}$$

By solving (4) we can obtain a_i, then submit it into (3) we have: $w = \sum_{i=1}^{n} a_i X_i Y_i$. As the quadratic programming problem satisfy the KKT condition, so we have $b = y_j - \sum_{i=1}^{n} y_i a_i^* K(X_i, X_j)$, with a_i^* is a coefficient larger than 0. As only when $a_i > 0$, it has effect on the value of Q_D. Therefore, we call the support vector corresponding to $a_i > 0$ as the support vector of X_i. Then we can get the decision function $f(X) = \text{sgn}(\sum_{i=1}^{n} y_i a_i^* K(X_i, X) + b)$

3. Rough Sets

Rough sets are proposed by Z.Pawlak in 1982, it is a data mining method which can be used to study the incompleteness of data, and uncertainty of knowledge. The basic idea of data reduction by using rough set theory can be outlined as follows: it find the decision regulation by the dependence relationship between the sample attribute and the decision attribute; then judge the importance by the degree of influence of attribute to the decision. By these procedures, the unimportant attribute can be removed, so as to achieve the classification ability of reduce the data characteristic and preserve the data nature.

Definition 1. Information system is a four number set I=<S,A,V,F>, with U is the nonempty sample set, and A is the attribute set, and V is the attribute value region, and F is the map, which can give a value from V for every sample attribute A in S.

For the training sample, there is some classification marks, such as the 42 dimensional intrusion sample of KDDCUP'99 is "normal", "abnormal" and so on. These attributes are called decision attributes. By introducing the decision attribute, we can obtain the decision graph by the information system.

Definition 2. The decision graph of information system is a four number set T=<S,A&{d}, V, F>, with A be the sample attribute, and its value a is called as condition attribute, and d is the decision attribute.

Definition 3. Indiscriminate relationship can be described as follows, in the decision graph DT, with $B \subseteq A$, for any sample in S, we have $F(a) = F(a^{'})$, then such a relationship is called the inseparable relationship between A and its subset B (B-indiscriminate relation), denoted by $IND_I(B)$, where $IND_I(B)$ refers to the indiscriminate relationship of attribute, i.e. the sample can not be discernible from attribute B. The decision indiscriminate relationship can be constructed based on the concept.

Definition 4. The indiscriminate relationship of decision is refer to the following fact, in $IND_I(B)$, we have $F(x,d) = F(x^{'},d)$, denoted by $IND_I(B,d)$.

Definition 5. The decision reduction refers to that in DT; we seek the smallest attribute set such that $IND_I(B,d) = IND_I(A,d)$ holds.

Though the decision reduction is a NP hard problem, there exist many fast reduction algorithms; this topic is beyond the discussion of this paper. Decision graph can be established by the decision graph discriminate matrix.

Definition 6. Suppose M is the decision graph discriminate matrix constructed based on DT, the element Mij on the (i, j) position is defined as follows,

$$M_{ij} = \begin{cases} \{a \mid a \in A \wedge f(x_i) \neq f(x_j)\} & f(x_i,d) \neq f(x_j,d) \\ 0 & f(x_i,d) = f(x_j,d) \end{cases}$$

By classifying the data protocol, and construct a decision graph for every group of data, then reduce the decision graph using the reduction algorithms, then we can obtain the different data set reduced from different data protocol.

4. The SVM intrusion detection method using protocol classification and rough Set

In the former investigation of rough set data reduction, the protocol is indiscriminate and the reduction is for all the data. There are two shortcomings in these approaches: firstly, all the data is strongly different structured, study the data using SVM, we need to introduce a new computation method for distance. On the other side, intrusion usually takes the leak of the different structured data. The indiscriminate protocol is just a broad detection method, it does not consider the different characteristic in different data, and hence these methods are not aimed. We propose the SVM intrusion detection method using protocol classification and rough sets, it is able to remove the shortcomings in the original methods, and is able to improve the detection time and the accuracy.

Classifying the protocol, using the rough set to reduce the data, then do training to the reduced data, i.e. the corresponding SVM input. The obtained training model is the SVM detector corresponding to different protocol. The SVM intrusion detection method using protocol classification and rough sets can be described as the following Fig 1.

In Fig 1, the real line illustrates the training process, the training data is classified according to protocol. Three different kinds of intrusion data is divided, denoted by TCP, UDP, and ICMP. Then carrying out the rough sets study for these three kinds of intrusion data, the studying procedure is denoted by T, U, and I. The reduced characteristic after study is used as the SVM study input; on the other hand, the reduced regularization is stored as the pre-definition process, denoted by reduction T, reduction U and reduction I. Three SVM study apparatus will become three detector models after study; they are stored as three detectors T, U and I. In Figure 1, the dash line denotes the detection process of the test data. The test data first classified by the protocol, then the reduction procedures are started based on different protocol data, the reduced data is inputted into corresponding detector, and the test results come from the detector. The

SVM intrusion detection method using protocol classification and rough sets can be described as the following algorithm:

Step 1: Input the training data, start protocol classification, the data is divided into TCP, UDP, AND, ICMP according to different data protocol; and they are stored in database.

Step 2: Start the rough sets study machine, reduce three kinds of data separably,then obtain their own reduced characteristic set T, U and I. Then construct a SQL sentence based on the characteristic set, which is stored as the pre-definition process. Finally, the reduced training data is inputted into the corresponding SVM study machine.

Step 3: Start the SVM study machine T, U and I, then obtain their own decision function by study. $f(X) = \text{sgn}(\sum_{i=1}^{n} y_i a_i^* K(X_i, X) + b)$, stored as detector U, T, and I.

Step 4: For the input data X to be detect, first do protocol classification, then start the pre-defined rough sets reduction process according to classification.

Step 4: Input the reduced data into the corresponding SVM detector, the output the detection results through the SVM detector, normal is denoted by +1, and abnormal is denoted by -1.

5. Experiment

5.1 The tested data

KDDCUP'99 is obtained in the real net work. It can be used to simulate the 5 classes including 23 different kinds of data arising from attack, these data can be used as experimental data in data mining. The 10% subset of the data has 494021 records, and each record has 41 characteristics, which incorporate the continuous, discrete and text data. We can put a note at the end of each record to show whether the data is normal. Therefore, such kind of data set is a classical multi-protocol multi-attack

different structured data set. By classifying the protocol for the normal and attack situations, the results are illustrated in Figure 2 as follows

Statistical results show that TCP protocol records are 190064, and ICMP protocol records are 283602, and UDP protocol records are 20354. In the TCP protocol classification, there are all different kinds of attack, and DoS attack most frequently. In the UDP and ICMP protocol, the R2L and U2L attack almost never appear. For the UDP protocol, the abnormal data includes DoS and Probe. For the ICMP protocol the DoS attack has 280 thousands records. The abnormal data is mainly DoS data.

After protocol classification, we begin to do test from selected training data and test data, the test results are outlined as follows,

(1). TCP test data: Choosing 30000 records from the TCP data set, where the normal data is 12802 items; and abnormal data is 17198 items (DOS has 16560 items, Prob has 422 items, R2L has 188 items, U2L has 8 items).

(2). UDP test data: Choosing 10173 records from the UDP data set, where the normal data has 9586 items, and abnormal data has 587 items (DoS has 489 items, Prob has 98 items).

(3). ICMP test data: Choosing 28353 records from the ICMP data set, where the normal data has 128 items, and abnormal data has 28225 items (DoS has 28105 items, and Prob has 120 items).

Taking 70% data randomly from the test data set for training; then leaving other 30% for test.

5.2 The reduction of the test data

Reducing the data by means of Rosetta tool Komorowski, 1997, P.403-407, and form different reduction set from the 41 reduced characteristic. The characteristic set reduced from TCP, UDP and ICMP are outlined in the following Figure 3, Figure 4 and Figure 5.

Choosing two groups of characteristic set, for example, take the first and the eighth from TCP, and take the sixth and the 30th from UDP, and take the sixth and the eighth from ICMP. By reducing the characteristic for the corresponding training data and test data, we can obtain the training data and the test data after characteristic reduction. Compare with the characteristic with the ones given in Sung P.405-411, we can see our approach has less characteristic and easier to deal with, and finally the test result shows that the our method can preserve high accuracy and much faster.

5.3 Data training and detection

In the test, we choose RBF function $f(x_i, x_j) = \exp(-(x_i - x_j)/2\sigma^2)$ as the SVM kernel, and adopt 5-Fold Cross Validation, embedded in the LibSVM software by Chihjen. The test is in three steps, firstly, we use grid search (grid.py command) to compute the optimal punish parameter C and σ^2, then obtain the training model by train the training data

using the optimal parameter. and finally test using the trained data. Take the example using 21000 TCP training data and 9000 test data, the parameter search is outlined in Figure 6. The optimal value is $C = 512, \sigma^2 = 0.03125$. By using these two parameters to train the 21000 TCP data, we obtain train.txt.model. Then we use this model to do test for these 90000 data. Finally, we obtain the training time, the detection time, and the accuracy.

For comparison reasons, the intrusion data and detection is divided into three situations. The first is to do test on the classified data by the complete characteristic. The second is to do test on the classified data by the reduced characteristic. The third is to do test on the unclassified data. The final test results are outlined in Figure 1, Figure 2 and Figure 3.

Comparing Figure 2 and Figure 3, we can discover that the training time and the detection time is shorten by using protocol classification, moreover, the detection accuracy is not damaged.

Comparing Figure 1 and Figure 2, we can see that using characteristic reduction and not using characteristic reduction has similar accuracy, however, the detection time and training time is saved obviously by using the characteristic reduction. Therefore, our conclusion is as follows, protocol classification along with characteristic reduction need less time, while using the complete characteristic need much more time, further more time is needed if protocol classification and characteristic reduction are all not carried out.

6. Concluding remarks

In this paper, we propose to use internet protocol classifying the intrusion data, and use rough sets to reduce unclassified data, and then do training for the reduced data, and finally produce a training model. In the test procedure, we first do protocol classification for the data, then do test for the model after training. We do some tests on the KDDCUP'09 data under three cases, the test results show that the new method produce more accuracy results, and need less training and test time. By theoretical analysis, the reason is as follows: as we have adopted protocol classification, which eliminate the difficulty caused by the unstructured protocol character, this reduces the time needed in dealing with data. On the other hand, as intrusion is due to the hole of protocol, so it is more targeted and the accuracy is not damaged because of the characteristic decrease. The future work will be implement a intrusion detection system based on the algorithm proposed in this paper. This will not only consider the protocol classification, but also need to consider that real internet intrusion actually a unsupervised character classification. Furthermore, it also needs multi-class classification technique research.

References

[DB/OL].http://www.csic.ntu.edu.tw/~cjlin/papers/libsvm.pdf.

Burges C. (1998). A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discover.* No. 2 .P. 121-167.

Chihjen L. LIBSVM: a library for SVMs (Version 2.6)

Frohlich H., Chapelle O., & Scholkopf B. (2003). Feature selection for support vector machines by means of genetic algorithm. *In: Proceedings of 15th IEEE International Conference on Tools with Artificial Intelligence.* No.3-5. P. 142 – 148.

http://kdd.ics.uci.edu/databases/kddcup99/task.htm.

Komorowski J.O., & ROSETTA. (1997). A rough set toolkit for analysis of data. *Fifth International Workshop on Rough Sets and Soft Computing.* Tokyo,Japan. P. 403-407.

Lin Y., &.Cunningham A. A New Approach to Fuzzy-Neural System Modeling. *IEEE Transactions on Fuzzy System*s, No.3. P.190-198.

Mukkamala S., Janoski G., & Sung H.(2003). Intrusion Detection Using Neural Networks and Support Vector Machines. *Proceedings of IEEE International Joint Conference on Neural Networks,* P.1702-1707.

Pawlak Z. (1982). Rough sets. International Journal of Information and Computer Sciences. No.11, P.341-356.

Roberto P., Guofei, G., & Wenke L. (2006). Using an Ensemble of One-Class SVM Classifiers to Harden Payload-based Anomaly Detection Systems. *In Proceedings of the Sixth International Conference on Data Mining (ICDM'06).*

Sung A. (1998). Ranking Importance of Input Parameters of Neural Networks. *Expert Systems with Applications*. No. 15. P.405-41.

Vapnik V. (1995). The nature of statistical learning theory. *New York: Springer-Verlag.*

Table 1. Experimental result with protocol difference but without reduction

protocal	best c, σ 2	Training time	Test time	Test correct rate
TCP	c=512, σ 2=0.03125	5.1s	1.6s	99.7889%
UDP	c=128, σ 2 =0.03125	2.3s	1.3s	99.9344%
ICMP	c=8.00, σ 2=0.078125	2.7s	1.4s	99.9765%

Table 2. Experimental results with protocol difference and reduction

protocal	Feature set	Best c, σ 2	Training time	Test time	Test correct rate
TCP	1	c=2048, σ 2=0.5	4.8	0.9	99.7287%
	8	c=32768, σ 2=0.125	4.2	1.2	99.765%
UDP	6	c=2048, σ 2=0.5	1.6	0.8	99.8689%
	30	c=2048, σ 2=8.0	1.8	0.9	99.9017%
ICMP	6	c=512, σ 2=5.0	2.1	1.0	99.9882%
	8	c=32, σ 2=0.078125	2.4	0.8	99.9765%

Table 3. Experimental results without protocol difference and reduction

Data set	best c, σ 2	Training time	Test time	Test correct rate
30000	c=512, σ 2=0.5	8.6	6.2	99.5%
10173	c=128, σ 2=0.125	4.7	3.7	99.8%
28353	c=64, σ 2=0.125	5.3	5.2	98.6%

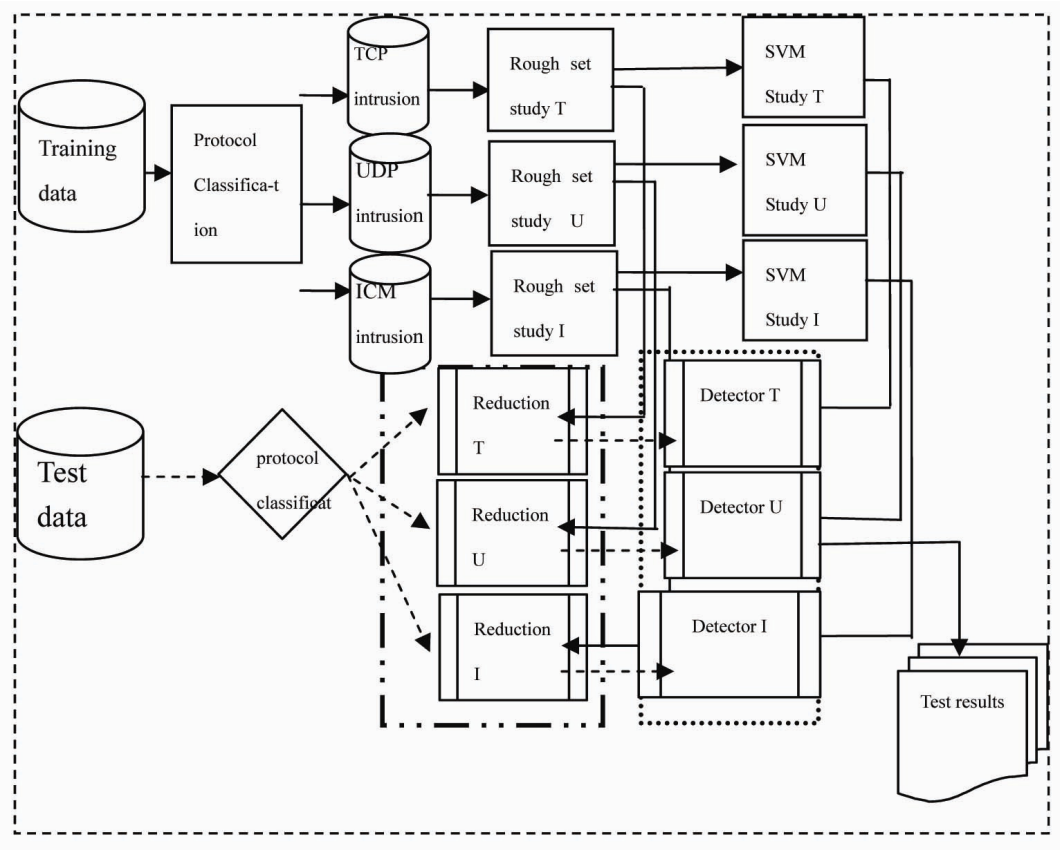

Figure 1. The SVM intrusion detection method using protocol classification and rough sets

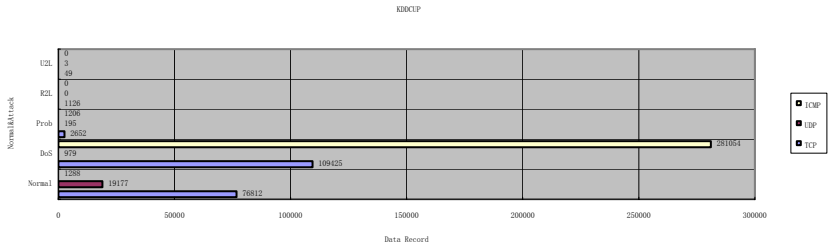

Figure 2. KDDCUP'99 intrusion data protocol classification.

NO	Feature set after Reduction	support	length
1	3,4,6,24,23,24,27,28,31,32,33,36,38	100	13
2	3,4,6,18,23,24, 27,28,31,32,33,36,38	100	13
3	3,4,6,14,23,24, 27,28,31,32,33,36,39	100	13
4	3,4,6,23,24, 27,28,31,32,33,35,36,39	100	13
5	3,4,6,23,24, 27,28,31,32,33,35,36,38	100	13
6	3,4,6,18,23, 24,27,28,31,32,33,36,39	100	13
7	3,4,6,10,23, 24,27,28,31,33,35,36,37,39	100	14
8	3,4,6,23, 24,27,28,31,32,34,35,36,37,39	100	14
9	3,4,6,10,23, 24,27,28,31,33,35,36,37,38	100	14
10	3,4,6,10,12,18,23, 24,27,28,31,32,34,35,36,37,38	100	17
11	3,4,6,10,12,14,23, 24,27,28,31,32,34,35,36,37,38	100	17

Figure 3. TCP reduced characteristic.

NO	Feature set after Reduction	support	length
1	3,5	100	2
2	5,24,34	100	3
3	5,6,36	100	3
4	5,31,36	100	3
5	5,34,35	100	3
6	5,33,36	100	3
7	5,32,33	100	3
8	5,6,32	100	3
9	5,8,36	100	3
10	5,32,34	100	3
11	5,8,32	100	3
12	5,23,33	100	3
13	5,30,34	100	3
14	5,31,32	100	3
15	5,34,36	100	3
16	5,33,34	100	3
17	5,29,34	100	3
18	5,8,29,35	100	4
19	5,8,23,35	100	4
20	1,5,23,34	100	4
21	5,6,30,35	100	4
22	3,33,35,36	100	4
23	5,23,34,40	100	4
24	5,6,29,35	100	4
25	5,8,30,35	100	4
26	5,8,30,33	100	4
27	5,23,31,34	100	4
28	5,24,30,33	100	4
29	5,29,31,33	100	4
30	5,6,23,34	100	4
31	5,30,31,33	100	4
32	5,33,35,36	100	4
33	5,30,31,35	100	4
34	5,6,29,31	100	4
35	5,29,31,35	100	4
36	5,24,33,35	100	4
37	5,24,29,33	100	4

Figure 4. UDP reduced characteristic.

NO	Feature set after Reduction	support	length
1	5,32	100	2
2	5,33	100	2
3	3,23,32,33	100	4
4	8,23,32,33	100	4
5	3,24,32,33	100	4
6	3,24,32,33	100	4
7	8,24,32,33	100	4
8	8,24,33,36,37	100	5
9	8,24,33,34,37	100	5

Figure 5. ICMP reduced characteristic.

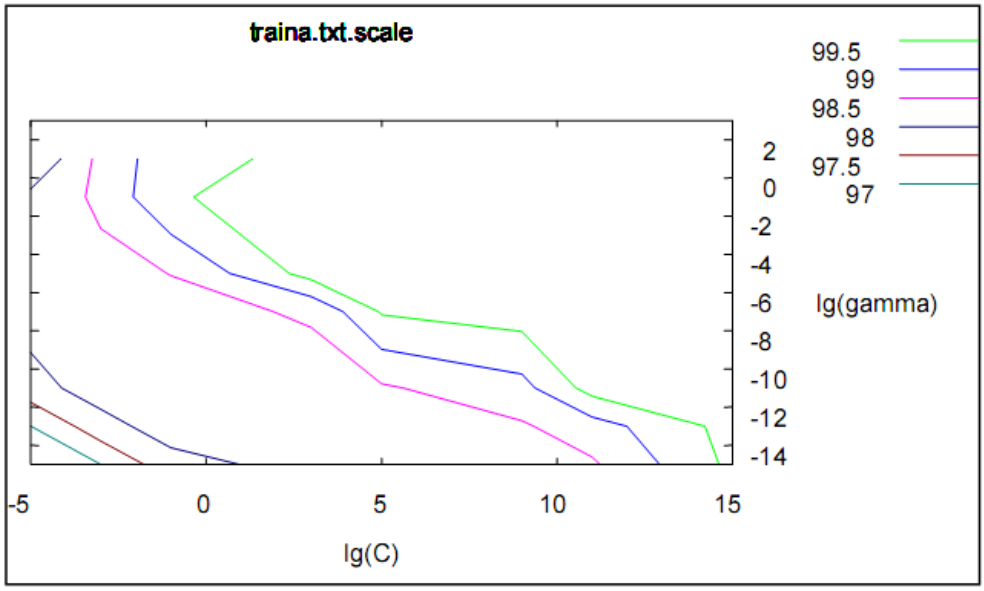

Figure 6. Parameter search of TCP training data

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 2, No. 4

November 2009

Framework for Interrogative Knowledge Identification

Fatimah Sidi (Corresponding author)

Computer Science Department, Faculty Computer Science & Information Technology

University Putra Malaysia

43400 UPM Serdang, Selangor, Malaysia

Tel: 60-12-203-8131 E-mail: fatimah@putra.upm.edu.my

Marzanah A. Jabar, Mohd Hasan Selamat, Abdul Azim Abd Ghani & Md Nasir Sulaiman

Faculty Computer Science & Information Technology, University Putra Malaysia

43400 UPM Serdang, Selangor, Malaysia

Tel: 60-3-8946-6555 E-mail: {marzanah, hasan, azim, nasir}@fsktm.upm.edu.my

Abstract

The difficulty of defining and capitalizing the knowledge in an organization from the business data captured in text files. These text files defined as unstructured document that is without a specific format example, plain text. Hence, this paper presents an Interrogative Knowledge Identification framework to identify unstructured documents that encompassed knowledge, information, and data. It tries to identify some high-level problems of the area from a higher perspective and then propose a possible solution thru the description of the framework. This research is an experimental approach using an appropriate test collection of unstructured documents. A system was developed based on the Interrogative Knowledge Identification framework. The results obtained are measured in terms of percentage of quantitative retrieval performance recall and precision metrics compared with an expert. This is to improve better understanding the process of making sense the information or knowledge residing in unstructured documents.

Keywords: Knowledge identification, Interrogative, Unstructured documents

1. Introduction

The difficulty of defining knowledge in unstructured documents is due to the paradox that knowledge resides in a person's mind and at the same time, it has to be captured, stored, and reported. For that, philosophers classify knowledge into knowing-that and knowing-how. Knowing-that is factual where data are stored in databases and facts can be recalled, processed, and disseminated. While knowing-how is actionable to do something, turning data into information and in turn into knowledge (Spiegler, 2003).

It is estimated that 90% of electronically available material is unstructured and the amount of unstructured textual documents, accessible through the web, intranets, news groups, etc. is enormously increased every year (Iiritano & Ruffolo, 2001). Hence, huge amount of unstructured documents are available on the web and intranets. The amount of information available to us is constantly increasing and our ability to absorb and process this information remains constant. Apparently, knowledge exists and is found everywhere (ubiquitous) in unstructured documents (Feldman, 1999), so identifying and extracting knowledge in unstructured documents is essential.

2. Unstructured Documents

A document is a paper or set of papers with written or printed information, especially of an official type. It is categorized into two classes, unstructured and structured. Unstructured document (a "flat" document) will not have any attributes. These types of documents usually have a title, but after that the content is not organized in any structured fashion examples news and scientific papers.

Structured documents have a well-defined hierarchical structure, such as titles and sections clearly marked with single or multiple level headings. Other attributes that create hierarchy, such as distinctive colour, underlines, boldness, etc., are also considered. Structured form/scheme is the way in which data or information are arranged or organized in rows

and columns.

Unstructured documents cannot be queried in simple ways. Therefore, knowledge contained in unstructured documents can neither be used by automatic systems nor could be understood easily and clearly by humans. Hence, identifying knowledge from unstructured documents to be easily realized and understood by humans is one of the most valuable areas to be explored.

3. Spectrum of Data, Information and Knowledge

There are three theories of knowledge (As-Sadr, 1987; Cornford, 1957). First, the idealistic notion like Plato believes that knowledge was a function of the recollection of previous information. Second, the materialistic notion that believes in five senses. They consider sense perception as the source or means of knowledge. Third, the Islamic notion that believes in the existence of matter as well as soul. By that, knowledge is a complex concept and it is not easy to define because it is not easily understood, perceived, and measured. It has absolute truth, or ground truth, which describes the rich truths of real situation experiences (Davenport & Prusak, 2000; Drucker, 2001).

However, most people have some understanding of what knowledge is. Knowledge, information, and data are not interchangeable concepts. A brief comparison of data, information, and knowledge based on literature are tabulated in Table 1. It shows that data, in and of itself is a symbol, are out of context and with no value until processed into useful forms. By adding meaning, values, and searching for context to make sense of data, this context reveals the structure or relationship (or both) that organizes the data into information. Knowledge is the process of making sense of information. Examples of knowledge are patents, recipes, formulas, instructions, and designs. Without the dimension of context, culture, tacit, and time, knowledge will be little more than information. Thus, knowledge has more to do with who is interpreting the information (their own principles and values) than the objective information on which it is based.

3. Theoretical Background

This section discusses theoretical foundation of the framework proposed. Philosophers see knowledge as justified true belief, while scientists see knowledge as documents empirical research, supported by Quigley and Debons, (1999). The word data, information, and knowledge have many meanings in many contexts, which are often embedded in documents, repositories, processes, practices, and norms. Therefore as a foundation of this research, the basis approach of knowledge understanding adopted is the scientist views. While, on the knowledge management (KM) understanding, it is concerned on the knowledge growth in an organization (Steels, 1993), such as organizing of knowledge (Gurteen, 1999; MingYu, 2002). Hence, it is important to facilitate knowledge transfer or discovery in unstructured documents.

Unstructured documents are stored at any time in history. These texts are stored in hardware and retrieved through software, which contains data, information, and knowledge, each with its own characteristics and value. According to Quigley and Debons (1999), a cognitive spectrum of data, information, and knowledge focus on data-as-thing, information-as-thing and knowledge-as-thing located within text strings. They reported an interrogative-based approach to differentiate and quantify information and knowledge within text. The interrogative-based approach is described as the "who, when, what, where, how and why" analysis. Analysis using interrogative theory makes distinctions between data, information, and knowledge as follows:

- Knowledge text that answers how/why in the problem space
- Information text that answers when/where/who/what in the problem space
- Data text that answers no question in the problem space

They reported their finding that parsing of the paragraph into interrogative strings yields consistent results and a quantification of information and knowledge within the text. Based on the perspectives above, that data, information, and knowledge focus as-thing located within text strings. It is recommended that elements of personal components and dimension of context, culture, tacit, and time should be included in the discussion of the Interrogative Theory. This is because there is a lack of fluid mix of framed experience, values, contextual information, and expert insight in the spectrum of data, information, and knowledge. Values and beliefs are integral to knowledge, determine a large part of what the knower sees, absorbs and concludes from his observations. People with different values "see" different things in the same situation and organize their knowledge by their values. By that, challenges to incorporate the personal components of values and beliefs could be seen as the gap in the discussion of the Interrogative Theory as this theory sees data, information, and knowledge only as-things. Therefore, a new perspective of looking upon the spectrum of data, information, and knowledge can be derived by unifying Interrogative Theory and personal components of values and beliefs.

4. Interrogative Knowledge Identification Framework

The interrogative knowledge identification is used to address the need for the mechanism to identify knowledge from unstructured document in order to extract them. Briefly restating the interrogative knowledge identification, it identifies the type of document by separation of text into knowledge, information or data and unifying it with personal

components of values and beliefs. The approach of answering interrogatively is used to answer the question within the text in unstructured document to identify knowledge.

Another important aspect is to understand the process of making sense the information that resides in the unstructured documents into knowledge. Knowledge must have enough characteristics of information in terms of its meaning, values and context to reveal its structure or relationship or both. Lack of ways or methods to organize information of unstructured document would produce different knowledge from the same piece of information in different brains. Barachini (2003) reports that the more contexts stored with a chunk of information, the better the interpretation and transfer of knowledge. Hence, introducing interrogative contextual information by adding more contexts to the information, organizing, and structuring them into interrogatively structured form will increase better understanding and interpretation of knowledge that resides in unstructured document.

The interrogative contextual information is derived from the incorporation of context and additional information annotation with context key facility. Context is an abstraction of the context factors, which are represented as concepts (Schilit & Theimer, 1994). It is further exploited by Lamming and Newman (1992) as contextual information, where information entered into the computer is tagged with context keys facilitating future retrieval by using those keys. It is any information that can be used to characterize the situation of an entity (Abowd, Dey, & Abowd, 1999). An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves. For that, the interrogative contextual information is utilized to understand the process of making sense of information into knowledge and maintain the meaning of the information. This is to gain the interpretation of the identical knowledge by classifying the main point of the unstructured document interrogatively.

The unification of the interrogative knowledge identification and contextual information with incorporation of personal components of values and beliefs is illustrated in Figure 1. The incorporation of personal components is motivated by looking at gaps and contradictions existing in retrieving documents through Internet by different people, culture, and values. Hence, the issue addresses is: how to identify knowledge in order to extract them in unstructured document? Different knowledge is produced from the same piece of information in different brains. Rationalization of the incorporation of personal components towards the interrogative knowledge identification is as follows:

- Nonaka (1994), personal components have a powerful impact on organizational knowledge;
- Davenport and Prusak (2000), knowledge is a fluid mix of frame experience, values, contextual information, and expert insight. It provides a framework for evaluating information. It originates in the mind of the knower to determine a large part of what the knower sees, absorbs, and concludes from his observations;
- Barachini (2003), knowledge is a private and personal thing. It is intuitive and strongly linked to the user's values and beliefs; and
- Virk (2004), manually transforming documents. Values are embedded because humans read documents, extract the values of existing fields, and then enter the values into a user interface.

The unification of the interrogative knowledge identification and contextual information with incorporation of personal components of values and beliefs as depicted in Figure 1 provides a proposal to establish an approach on transformation of extracted knowledge in unstructured documents by identifying, organizing, and structuring them into interrogative structured form. It is used to transform information in unstructured documents into knowledge. It is also used to understand the process of making sense of information into knowledge; maintain the meaning of the information; and gain the interpretation of the identical knowledge by classifying the main point of the unstructured documents interrogatively. It is designed to ease the burden of work, through augmentation and automation, allowing resources to be applied efficiently to the tasks for which they are most suited. It is important to note that not all knowledge extractor are computer-based, as paper and pen can certainly be utilized to generate, codify, and transfer knowledge (Ruggles, 1997). For the purpose of this research, however, the tools covered are primarily the technological ones due to their quick evolution, dynamic capabilities, and organizational impacts.

5. Research Setting

The research setting involves the development of the system based on architecture of the proposed framework; i.e. Malay*IK*-Ontology (Malay Interrogative Knowledge Ontology). Basically, the system consists of these four processes:

i. Prepare the unstructured documents to be processed and converted it into extension of plain text file.

ii. Invoke lexicon identifier that uses lexicon interrogative analysis matching rules. It is used to identify and extract knowledge in each of the complete sentences written in the unstructured document. It is also used to extract interrogative lexical constructs from the individual unstructured document.

iii. Invoke object recognizer that uses matching rules of object interrogative analysis to extract ontological constructs from the interrogative lexical constructs. It is used to populate objects and map the objects with ontology engineering. It

is a mechanism of a knowledge structure to represent the concept and relationship of the abstract model on how people think about things in the world.

iv. Transform ontological constructs to populate database scheme by connecting ontology model with conceptual modeling of object-relationship model. This is used to structure the extracted knowledge into interrogative structured form.

From the above processes, it can be simplified as shown in Figure 2.

This research is an experimental approach research using the Malay language. Therefore, an appropriate Malay test collection of Malay unstructured documents is required. Different topics are drawn such as main news, technology, editorial columns, sports, letters, and e-mails, while texts from children story books, articles, and magazines are drawn from Internet or retyped from the printed materials. This is a stratified population of data samples. In order to guarantee equal representation of each identified strata, a stratified random sampling is used. It is based on the number of words in the unstructured document and text which cover simple sentences constructed in Malay language. Each document drawn is assigned with a serial number and number of words.

The documents drawn are grouped according to the source of documents and range of number of words. The points of the range are defined at positions of 50-150, 151-300, 301-500, and the final range is more than 500. For each range, five unstructured documents are selected and sorted in ascending order by total number of words. The total number of words needed for Malay unstructured documents test collection is about 15% of 42,733 words from Malay Interrogative Knowledge Corpus (Malay*IK*-Corpus).

The Malay language corpus is derived from 6,000 word entries (about 4,000 root words and 2,000 derivations), a Malay language dictionary of Kamus Dewan published by Dewan Bahasa Perpustakaan (2005). It is also derived from other dictionaries of Kamus Imbuhan Bahasa Melayu (Ali, Shariff, & Dewa, 1993), Kamus Dwibahasa Oxford Fajar (Hawkins, 2001), and Kamus Komprehensif Bahasa Melayu (Othman, 2005). The sample used in this experiment, 15% of 42,733 words from Malay*IK*-Corpus are sufficient and justified to produce better results in extracting identified knowledge. It is more than the suggested by Gay and Airasian (2003, page 113) for sample of more than 5,000 units, a sample size of 400 (8%) should be adequate.

The results obtained are measured in terms of percentage of quantitative retrieval performance recall and precision metrics (Baeza-Yates & Ribeiro-Neto, 1999). The accuracy of the knowledge extracted is measured by precision (fraction of the retrieved knowledge which has been relevant), and recall (fraction of the relevant knowledge which has been retrieved). Comparison of results on the testing and analysis of the Malay*IK*-Ontology implemented is done with an expert evaluation. The Malay unstructured documents collection is given to the expert to identify the knowledge that resides in the collection interrogatively. The expert then validates the system generated output based on the interrogative criteria. The expert reffered to is Prof. Dr. Hj. Awang Sariyan from Academy of Malay Studies, Universiti Malaya. He is also a member of the Language Commitee Organizer Board, Institute of Language and Literature, Malaysia.

6. Results and Discussion

The analysis of results confirmed by a significant accuracy in identifying and extracting knowledge by using interrogative element of why. Unfortunately, this is not true for the interrogative element of how. Both these interrogative elements are used to identify knowledge within the text in unstructured document. Moreover, the analysis of results has also confirmed significant accuracy in identifying and extracting information for the interrogative elements of what and who. Unfortunately, the accuracy differences are not significant for the interrogative elements of where and when. The reasons for the performances differences are possibly caused by the quality of various formats and styles of writing the Malay unstructured documents collection used.

7. Conclusion

The paper presents a development of a system based on architecture of the Interrogative Knowledge Identification framework to identify unstructured documents that encompassed knowledge, information, and data. It also improved better understanding the process of making sense of information into knowledge. This framework can be used to organize and structure knowledge and information into interrogatively structured form which increased better understanding and interpretation of knowledge that resides in unstructured document. It shows a clear knowledge organization and structuring concept that can increase understanding of the concept among the community. This leads to potential increase sharable and reusable of the concept among the community. Moreover, it can be used to facilitate student learning in understanding the information and knowledge resides in the unstructured document.

Our future work is to enable the incorporation of personal components of values and beliefs integrate and contextual information within the proposed framework. This is to maintain the meaning of the information and gaining the interpretation of the identical knowledge in unstructured document which facilitate identical knowledge perceived by different people.

References

Abowd, G. D., Dey, A. K., & Abowd, G. D. (1999, 27-29 September). Towards a Better Understanding of Context and Context-Awareness. *Paper presented at the Proceeding 1st International Symposium on Handheld and Ubiquitous Computing (HUC '99)*, Karlsruhe, Germany.

Ali, H. M., Shariff, M. N. M., & Dewa, W. M. W. (1993). Kamus Imbuhan Bahasa Melayu Edisi Kedua. *Kuala Lumpur, Malaysia: Penerbit Fajar Bakti Sdn. Bhd.*

As-Sadr, A. M. B. (1987). Our Philosophy. *USA: Routledge and Taylor & Francis Group*.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern Information Retrieval. *New York: Addison Wesley*.

Barachini, F. (2003). Frontiers for the Codification of Knowledge. *Journal of Information & Knowledge Management*, 2(1), 41-45.

Cornford, P. F. M. (1957). Plato's Theory of Knowledge. *Indianapolis, Indiana: Bobbs-Merrill Company, Inc.*

Davenport, T. H., & Prusak, L. (2000). Working Knowledge: How Organizations Manage What They Know*: Harvard Business School Press*.

Dewan Bahasa Perpustakaan. (2005). Kamus Dewan Edisi Keempat. *Kuala Lumpur, Malaysia: Dewan Bahasa dan Pustaka*.

Drucker, P. F. (2001). The essential Drucker: Selections from the management works of Peter F. Drucker*: New York: Harper Business*.

Feldman, R. (1999, August). Mining unstructured data. *Paper presented at the Tutorial notes of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*.

Gay, L. R., & Airasian, P. (2003). Educational Research: Competencies for Analysis and Application, Seventh Edition. *Merrill, New Jersey: Upper Saddle River*.

Gurteen, D. (1999, February). Creating a Knowledge Sharing Culture. *Knowledge Management Magazine 2*.

Hawkins, J. M. (2001). Kamus Dwibahasa Oxford Fajar Edisi Ketiga. *Kuala Lumpur, Malaysia: Penerbit Fajar Bakti Sdn. Bhd.*

Iiritano, S., & Ruffolo, M. (2001, 3-7 September). Managing the knowledge contained in electronic documents: a clustering method for text mining. *Paper presented at the Proceeding of the 12th International Workshop on Database and Expert Systems Applications*.

Kaipa, P. (2000). Knowledge architecture for the twenty-first century. *Behaviour & Information Technology*, 19(3), 153-161.

Lamming, M. G., & Newman, W. M. (1992). Activity-based information retrieval: Technology in support of personal memory. *Paper presented at the Proceeding of the IFIP 12th World Computer Congress on Personal Computers and Intelligent Systems - Information Processing '92*.

MingYu, C. (2002). Socialising Knowledge Management: The Influence Of The Opinion Leader. *Journal of Knowledge Management Practice*.

Nonaka, I. (1994). A Dynamic Theory of Organizational Knowledge Creation. *Organization Science*, 5(1), 14-37.

Othman, A. (2005). Kamus Komprehensif Bahasa Melayu. *Selangor Darul Ehsan, Malaysia: Penerbit Fajar Bakti Sdn. Bhd., a subsidiary of Oxford University Press*.

Quigley, E. J., & Debons, A. (1999). Interrogative theory of information and knowledge. *Paper presented at the Proceeding of the 1999 ACM SIGCPR conference on Computer personnel research*, New Orleans, Louisiana, United States.

Ruggles, R. (1997). Knowledge Tools: Using Technology to Manage Knowledge Better. *Paper presented at the Innovation Working Paper*, Ernst & Young Center for Business Innovation in Cambridge, Mass., and Business Intelligence Ltd. .

Schilit, B. N., & Theimer, M. M. (1994). Disseminating Active Map Information to Mobile Hosts. *IEEE Network*, 8(5), 22-32.

Spiegler, I. (2000). Knowledge Management: A New Idea or a Recycle Concept? *Communications of Association for Information Systems*, 3(14).

Spiegler, I. (2003). Technology and knowledge: bridging a "generating" gap. *Information & Management*, 40(6), 533-539.

Steels, L. (1993). Corporate knowledge management. *Paper presented at the Proceeding of ISMICK'93*, Compiègne,

France.

Virk, R. (2004). Transforming Unstructured Content into "Meaningful" XML. Retrieved 8 January, 2004, from http://www.dmreview.com/whitepaper/WID413.pdf

Table 1. Comparison of data, information and knowledge

Supporting Literatures	Data	Information	Knowledge
(Quigley & Debons, 1999)	- symbols, numbers, or characters	- process, informed mental state, commodity, product, or thing	- as a thing
(Davenport & Prusak, 2000)	- set of discrete, objective facts about events - structured records of transactions in organizational context	- document or audible or visible communication - has meaning the "relevance and purpose" - data becomes information when its creator adds meaning by adding value	- a fluid mix of framed experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information
(Spiegler, 2000; , 2003)	- record, store and maintain attributes	- when add value in some way	- when it adds insight, abstractive values, and better understanding
(Kaipa, 2000)	- symbols represent objects, events and/or their properties - out of context - no value until processed into useful forms	- a function of processed or structured data containing both the data and its relationship - provide objective descriptions - content oriented	- has both collective and personal components - has tacit and explicit nature - is the process of making sense of information

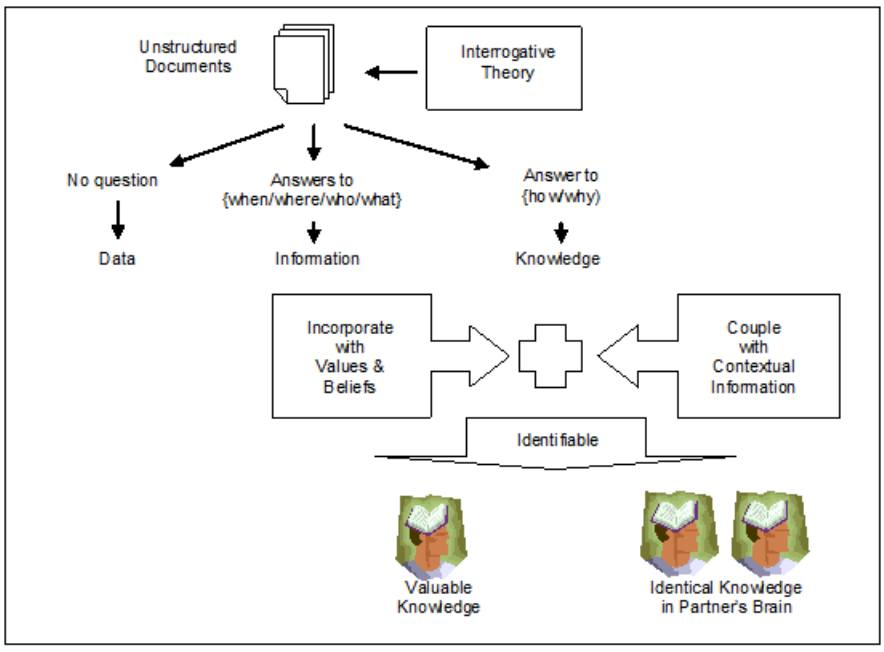

Figure 1. Interrogative Knowledge Identification Framework

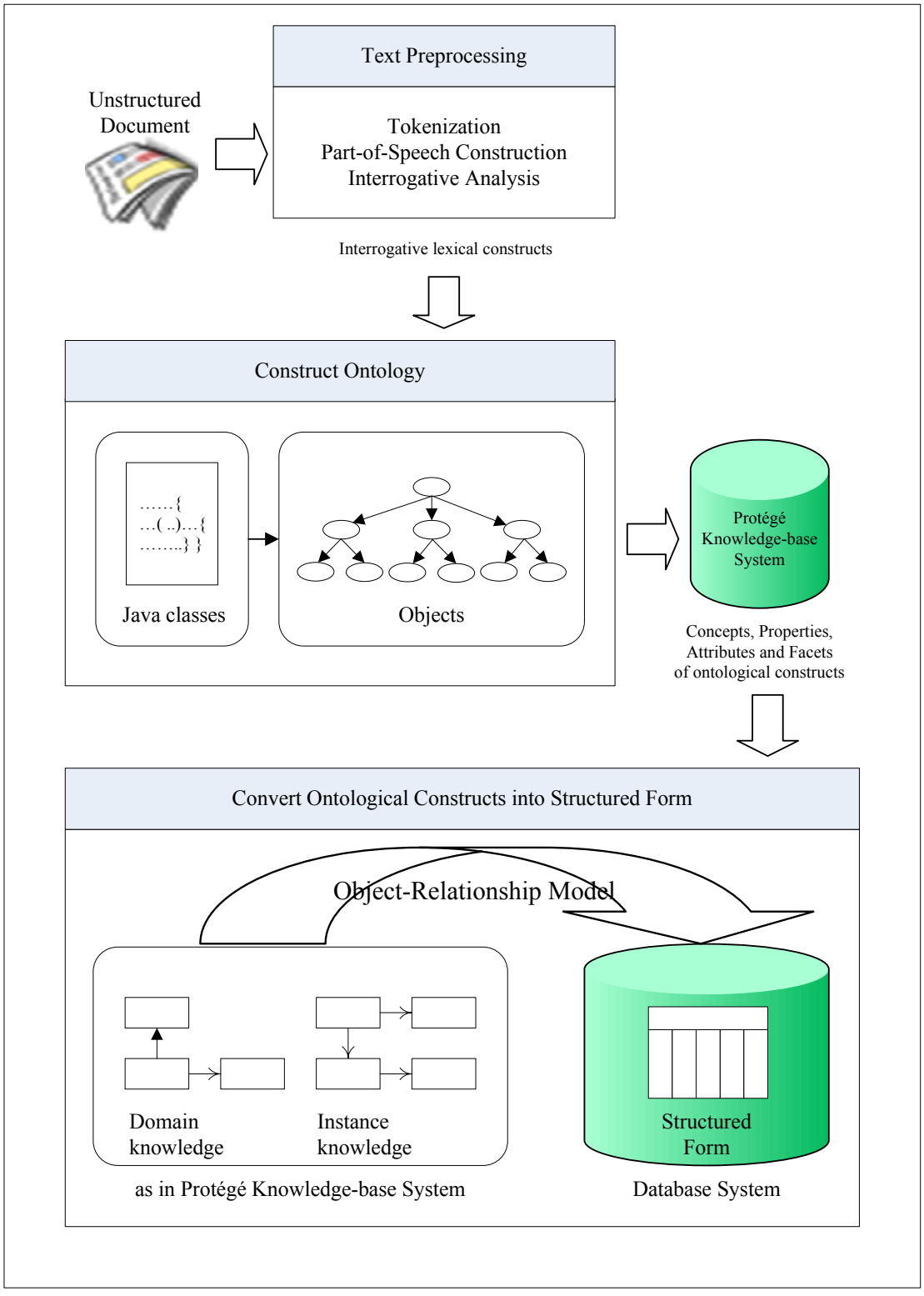

Figure 2. The Malay*IK*-Ontology Model

Computer and Information Science
www.ccsenet.org/journal.html
Vol. 2, No. 4
November 2009

Analysis and Design of ETL in Hospital Performance Appraisal System

Fengjuan Yang
Department of Computer and Information Science, Fujian University of Technology
Fuzhou 350100, China
Tel: 86-591-2270-2070 E-mail: yangfj99@126.com

Abstract

Taking the hospital performance appraisal items as the background, the method and mechanism of ETL process, data extraction, data cleaning, data transformation and increment updating are concretely analyzed and designed in the article. The data preparation area is added in the ETL process of the system, and the monitoring and restarting mechanism is set up in the system, which can effectively enhance the efficiency of ETL process.

Keywords: ETL, Hospital performance appraisal, Data Extraction, Data Transformation

1. Introduction

ETL (Extract-Transform-Load) is the process of data extracting, data transforming and data loading, and it is the core and soul of BI/DW (Business Intelligence/Data Warehouse), and it can integrate and enhance the value of data according to uniform rules, and it is responsible for the process transforming from the data source to objective data warehouse, and it is the important approach to implement data warehouse. The hospital performance appraisal system involves many business systems which are developed by different development teams in different terms. Because the data sources differ in thousands ways in data content, data format and data quality, which brings certain difficult and larger workload to the ETL process of the system. It is one of key factors to design a high-effective ETL process for constructing the system.

2. System structure of the performance appraisal system

The performance appraisal system is the appraisal system composed by KPI (Key Performance Indicators) which are independent and associated, and can completely express the appraisal requirements. KPI are the references for performance management, objective management, organization design and strategic management, and the performance appraisal method generally emphasized in modern enterprises (Dai, 2007, P.91).

The data sources of the hospital performance appraisal system root in multiple heterogeneous data sources such as HIS, human resource management system, OAS, financial management system and material management system, and the databases of these systems include Oracle and SQL Server. And the data sources with other document characters are also included in the hospital performance appraisal system. The hospital performance appraisal system is to utilize the mass data produced by the business support system, adopt the computer technologies such as data warehouse and data digging to extract, integrate, analyze and dig data, and provide timely, exact and scientific appraisal references for the performance appraisal of the hospital. The structure of the hospital performance appraisal system is seen in Figure 1.

3. Introduction of ETL tools

The representative ETL tools include Informatica, Datastage, OWB and Microsoft DTS at present.

Informatica Power Center is the advanced ETL tool in the industry, and it can conveniently extract data from heterogeneous system and data sources for users to establish, deploy and manage the data warehouse of the enterprise, and help the enterprise to make quick and exact decisions. This product can provide extensive supports for the application and data sources such as ERP system (Oracle, PeopleSoft and SAP), CRM (Siebel), electric business data (XML, MQ Series), legacy system and host computer data. The ETL tool in the solution project of IBM DB2 is Visual Warehousing which is included in Data Warehousing Manager. The Datastage of Ascential is the manufacturer with the highest share in the market, and it is the solution to support various systems and platforms such as host computer, ERP, UNIX and NT. The basic frame of Oracle Warehouse Builder includes two parts, i.e. design environment and operating environment. OWB can automatically generate the SQL codes corresponding with database object, and these codes can be distributed into the database, and ETL is implemented by the codes which are distributed into the database by the Oracle Enterprise Manager. DTS is the data integrating tool of Microsoft which can complete data extracting, transforming and loading on the Windows platform (Webmaster, 2006).

All above tools are all-purpose graph interface tools, and they can screen complex coding tasks, enhance the speed and reduce the difficulty. For the hospital performance appraisal system, the extracted data quantity is huge and the parameters are numerous. To enhance the efficiency and the flexible expansibility of the system, the ETL in the article combines OWB tool with SQL to quickly establish the ETL project by OWB, and utilize the flexibility of the SQL method to enhance the development speed and efficiency of ETL.

4. ETL design in the hospital performance appraisal system

Figure 2 is the structure of the ETL, and in the designing process of the system, the extracting program first extracts exterior data to the data preparation area, and then the system cleans the data in the data preparation area, and transforms the data according to the data model, and finally loads the transformed load to the data warehouse.

4.1 Data preparation area

Because the data extracting, cleaning and loading of the data warehouse need long work time, and to reduce the influence to the data source system and enhance the extracting efficiency, the system sets up the data preparation area.

The data preparation area is the work platform of data preparation, and its function mainly includes three aspects. First, the data extracted from the data source in the data preparation area can enhance the extracting efficiency and reduce the data extracting time, and reduce the influence of data extraction on the business support system. Second, the data preparation area can extract multiple data sources, and enhance the reliability and coherence of the extracted data. Third, some simple data preprocessing can be made in the data preparation area, which can enhance the efficiency of cleaning and transforming (Qi, 2005).

4.2 Restarting mechanism

Set up the restarting mechanism of data extracting, cleaning and loading in the data preparation area. In the data extracting, cleaning and loading process, because of the reasons of the system or some unpredictable factors, these activities will often fail, and if the system is restarted after failing, large numbers of resources of the system will be wasted. Therefore, the monitoring mechanism of data extracting, cleaning and loading in the data preparation area can be set up to dynamically monitor these activities, once they failure, the system can restart from the position of failing. To complete this mechanism, the data extracting, cleaning and loading activity can be divided into many approaches, and when the system enters into certain approach, it can hold the present status.

4.3 Data extraction

When designing the data extraction, the system should mainly consider the extracting mode, extracting content and increment updating of data.

4.3.1 Extracting mode

The data extracting mode includes the active mode and the passive mode. The active mode means that the source system actively extracts the data according to the data format defined by two parties. The active mode will make the source systems or other development teams depend on the performance and network of the source system. The passive mode means that the ETL program directly interviews the data source to acquire the data mode, and under this mode, the ETL works independently and extracts the data by itself.

The system in the article adopts the passive data extracting mode because the extracting time can be flexible and the structure change of the business system can not influence the normal work of ETL program.

4.3.2 Extracting content

The second problem of the data extracting is “what data the system extract”. The hospital performance appraisal system involves many tables, and the extracting must fulfill two conditions, and the first one is that the extracted data should fulfill the requirements of corresponding indexes in the performance appraisal system, and the second one is that the extracting process should not influence the performance of the original business system. Therefore, the system adopts the combination of the full extraction and the increment extraction. For the tables with small data quantity, the full extraction can be adopted, and all dictionary tables in the system all belong to small tables, and the data quantity is less than thousands of records. And these tables include the section office table, the doctor table, the medicine table and the illness table, and for these tables, the full extraction should be adopted. For the tables with large data quantity, such as the charge list, the illness diagnosis record table and the patient record table, and the data quantity of these tables can achieve ten-millions-class, so relatively flexible increment extracting modes must be adopted. The increment extraction can reduce the extracting data quantity, reduce the influences on the transformed and loaded data quantity, network flux and the business system performance, and enhance the performance of the whole process.

4.3.3 Updating of increment

The increment updating is the most important problem in the ETL process design, and extracting mode of data increment directly influences the performance of the system. At present, the changeable data methods in common use

include trigger, time-stamp and log comparison. The extracting mode of time-stamp is used in the fields with time-stamp, and it can distinguish whether the record belongs to the newly added record, and the comparison of the ending time of the last extracting and the time-stamp field in the table can decide the extraction of the data increment.

Taking the in-patient charge list increment of HIS as the example, according to the control flow table of the system ETL (Table 1), the extracting period is week, and the extracting time is the two o'clock in every Sunday, and the extracting SQL sentences are

//select * from in-patient charge list

where pricing date and time > to_date('05/10/2008 23:59:59', 'DD/MM/yyyy HH24:MI:SS')

and pricing date and time <=to_date('11/10/2008 23:59:59', 'DD/MM/yyyy HH24:MI:SS")"//

For the table without time-stamp fields, the log comparison increment extracting mode can be adopted to analyze the log of the database and judge the changing data. The CDC (Changed Data Capture) of Oracle is the representative technology in this aspect, and CDC can identify the changed data after last extracting. By means of CDC, when the source table implements many operations such as inserting, updating or deleting, the system can extract the data, and the changed data are stored in the changed table of the database. So the changed data can be captured, and are provided to the objective system by a kind of controllable mode through the database view.

4.4 Data cleaning and transformation

The data extracted from the business system are put into the data preparation area and cleaned in the data preparation area, and the dirty data can be filtrated. Then the system completes incomplete data and transforms the cleaned data according to the designing requirements.

4.4.1 Data cleaning

The data falling short of requirements mainly include incomplete data, false data and repetitive data.

(1) Incomplete data. These data are some information deficiencies such as the sex of patient, birth date and region. These data can be completed by the concealed information according to patients' ID number. For the data which can not be completed, some user-defined types can be used for later analysis, for example, when the patient's family address is deficient and can not be completed, fill in "undefined", and these data can be extracted in the future conveniently, and deleted according to actual situation.

(2) False data. The main reason of the false data is that the business system is not complete, and after the data are incepted, the data are directly wrote into the backstage database without being judged, for example, the numerical data follow an enter operation, or the input of the character string is false, the date format is not correct or the date is beyond the mark. The name of the doctor is "Zhang San", but the input may be "Zhang Shan" or "Zhang Sun". These data should be classified, and found out by the SQL sentence, and extracted when the business department modifies the business system.

(3) Repetitive data. For these data especially in the dimensional table, all fields recorded by repetitive data should be educed to confirm and process by the business department.

Data cleaning is a repetitive process, and it can not be completed in several days, and it can continually discover and solve problems. The business department will confirm whether the data need to be filtrated and modified, and the filtrated data are wrote into the data table by the form of Excel file, and in the initial stage of ETL development, the e-mails transmitting the filtrated data to the business department will make the department to modify the mistakes as soon as possible and regard the data as the references in the future. The data cleaning should validate each filtrated rule and be confirmed by the business department, and should not filtrate useful data.

4.4.2 Data transformation

The task of data transformation mainly includes the inconsistent data transformation, the transformation of data granularity, and the calculation and integration of some business rules.

(1) Inconsistent data transformation. This process is a process integrating data with same type in different business operations, for example, the sexes in HIS are denoted by "M and F", but they are denoted by "Male and Female" in the office system, and they are denoted by "0 and 1" in the financial system. After extracting, the sexes are denoted by "0 and 1" uniformly. If the quantity of the data needing transformation is large, the transformation comparison table can be designed to conveniently transform the data, for example, the section office codes include hundreds even thousands of records, and they can be designed as the comparison table such as Table 2 to convenient for the transformation of the section office codes.

(2) The transformation of data granularity. The business system generally stores fined data, but the data in the data warehouse are used for analysis, so generally the data in the business system will be integrated according to the data

warehouse granularity.

(3) The calculation of business rules. Different enterprises have different business rules and different data indexes, and these indexes sometimes can not depend on simple adding operations, and these data indexes calculated in ETL need to be stored in the data warehouse for analysis and use.

4.5 Data loading

Data loading is to load the data extracting, transforming and cleaning in the source business system into the data warehouse. In the system loading process, not only the data loading method should be considered from the performance angle, but also the data validating mechanism should be established, for example, validating the input and actual loading record amount, and processing and transferring the abnormal mistakes. This system adopts asynchronous and batch processing mode to load the data, and because the data loading involves many system resources, and needs the processing, interior memory and exterior memory equipments of data source and data warehouse. The data loading of data warehouse is implemented at the two o'clock, because the business system in this period is spare.

In addition, the loading and renovating of large number of data can only be implemented in the first-time data loading when the data warehouse is just established, and the sequent data loading always needs adopting increment data loading method. When implementing increment data loading, some necessary preparation works should be completed in the loading in the data preparation area.

5. Conclusions

The effective strategies and implementation projects are proposed in the article for the data extraction mode, data cleaning, data transformation and data loading mode in the ETL design process of the hospital performance appraisal system, and these strategies and projects can ensure the clarity and high-efficiency of the whole ETL process, enhance the veracity and reliability of data in the system, and provide powerful data guarantee for the data integration and data digging with high quality for the hospital performance appraisal system. The system proposed in the article has been implemented successfully and acquires good effects in certain provincial Class Ⅲ-A general hospital.

References

Dai, Huazhen, He, Liangyu & Yang, Feihong. (2007). Design and Realization of Bank Achievement Inspection System Based on ETL Technology. *Modern Computer*. No.273(12). P.91.

Qi, Yinfeng & Wang, Manshu et al. (2005). Research of Chinese Enterprise Investment and Financing Behaviors: Based on Results of Questionnaires. *Management World*. No.3.

Webmaster. (2006). White Book of Disaster Recovery. [Online] Available: http://www.ibm.com.

Table 1. ETL control flow

No.	Data source	Name	Extracting mode	Extracting condition	Operating time	Sign of success
1	HIS	Section office dictionary table	Full		2008.10.12 02:00:00	Success
2	HIS	In-hospital charge list	Increment	2008.10.5 23:59:59	2008.10.12 02:00:02	Success
3	Office system	Human resource table	Full		2008.10.12 02:10:00	Success
4	Financial system	Salary table	Increment	2008.10.5 23:59:59	2008.10.12 02:10:05	Success

Table 2. Comparison table of section office codes

HIS system code	Name	Corresponding	Code of section office	Name of section office
C210	Earthquake relief nursing unit		FF	Flexible section office
C308	ENT nursing unit		FF02	ENT nursing unit
C30801	ENT medicine-chest		FF0201	ENT medicine-chest
A101	Director of hospital		1	Director of hospital
A102	Department of medical affairs		2	Department of medical affairs
A103	Department of politics		3	Department of politics
A104	Department of hospital affairs		4	Department of hospital affairs
A105	Department of nursing		5	Department of nursing
A10208	Department of planning		6	Department of planning
A10201	Department of medical treatment		7	Department of medical treatment
A10202	Departmnet of training		8	Departmnet of training
A10203	Department of economic management		9	Department of economic management
A1020301	Office of outpatient service and charge		901	Office of outpatient service and charge
A1020302	Office of plastic surgery department charge		902	Office of plastic surgery department charge

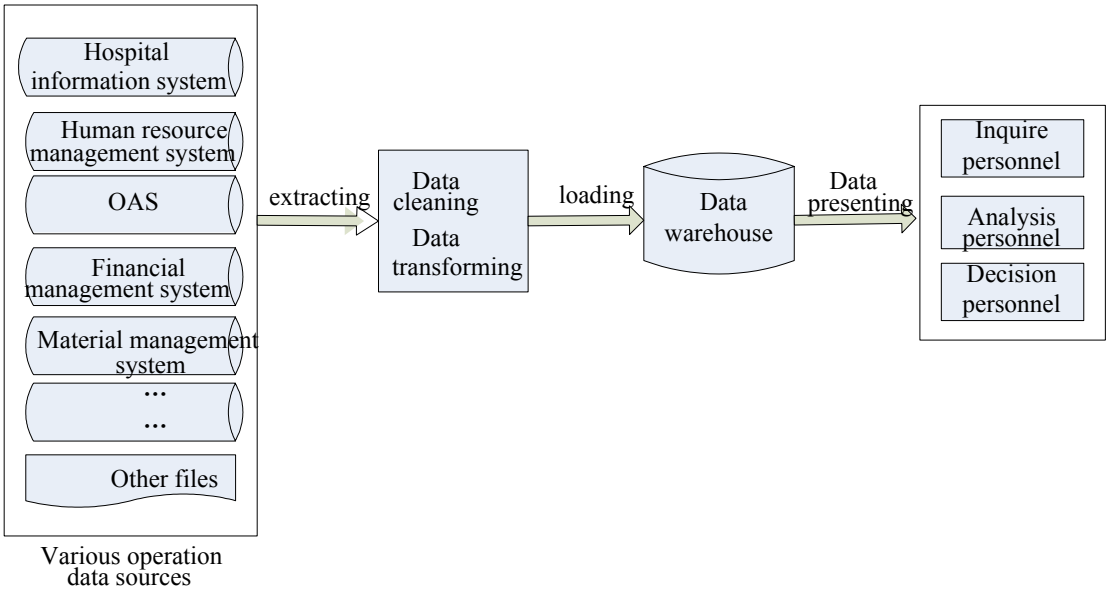

Figure 1. Structure of Hospital Performance Appraisal System

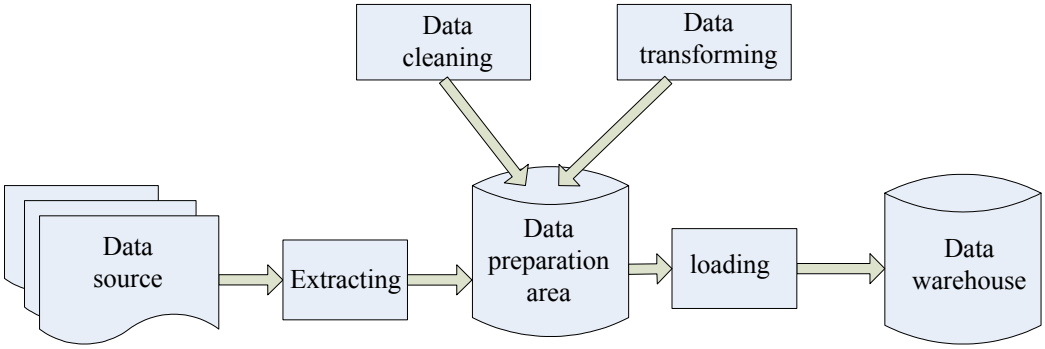

Figure 2. ETL System Structure

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 2, No. 4

November 2009

The Use of ICT in Public and Private Institutions of Higher Learning, Malaysia

Siti Rafidah Muhamat Dawam
Faculty of Computer Sciences and Mathematics, Universiti Teknologi MARA Kedah
P.O.Box 187, 08400 Merbok, Kedah, Malaysia
Tel: 60-4-456-2462 E-mail: srafidah192@kedah.uitm.edu.my

Khairul Adilah Ahmad
Faculty of Computer Sciences and Mathematics, Universiti Teknologi MARA Kedah
P.O.Box 187, 08400 Merbok, Kedah, Malaysia
Tel: 60-4-456-2450 E-mail: adilah475@kedah.uitm.edu.my

Kamaruzaman Jusoff (Corresponding author)
Faculty of Forestry, Universiti Pertanian Malaysia, 43000 Serdang, Selangor, Malaysia
Tel: 60-3-8946-7176 E-mail: kjusoff@yahoo.com

Taniza Tajuddian
Faculty of Computer Sciences and Mathematics, Universiti Teknologi MARA Kedah
P.O.Box 187, 08400 Merbok, Kedah, Malaysia
Tel: 60-4-456-2461 E-mail: taniza@kedah.uitm.edu.my

Shamsul Jamel Elias
Faculty of Computer Sciences and Mathematics, Universiti Teknologi MARA Kedah
P.O.Box 187, 08400 Merbok, Kedah, Malaysia
Tel: 60-4-456-2181 E-mail: shamsulje@kedah.uitm.edu.my

Suhardi Wan Mansor
English Language Department, Universiti Teknologi MARA Kedah
P.O.Box 187, 08400 Merbok, Kedah, Malaysia
Tel: 60-4-456-2190 E-mail: suhardiwm@kedah.uitm.edu.my

Abstract

This study examines the extent of ICT utilization among the members of Faculty A of four public higher learning institutions (IPTA) and seven private higher learning institutions (IPTS) in Northern Malaysia. Its focus is on a) to investigate the extent of ICT resources provided by universities authorities, b) focus on types and extent of ICT usage in daily activities, c) to explore the ICT proficiencies level and d) to investigate the level of ICT integration in teaching activities. A total of 76 responses out of 77 from IPTA and only 105 out of 108 responses of IPTS are usable for further analysis in this study. Findings indicate that in the IPTA, though the facilities provided are not as plenty as in IPTS, the

level of usage is quite encouraging. While in the IPTS, the levels of ICT usage among the educators are still not satisfactorily. Results also indicated that usage frequencies are more prone on informative in nature, besides integrating computer technology. Furthermore, the study also indicates that there were considerable differences in the use of ICT by educators in their perceived proficiencies and integrating computer technology. This study could be improved by expanding the total sampling population to all faculties in both universities. Methods of analysis could also be varied beyond the descriptive analysis done. Factors that could hinder the level of ICT usage by the educators could also be studied.

Keywords: ICT, Usage, Resources, Frequencies, Proficiencies, Integration

1. Introduction

The first computer system in Malaysia was implemented in 1996 (Chan, 2002). Since then, the Government has introduced various initiatives to facilitate the greater adoption and diffusion of ICT to improve capacities in every field of business, industry, education, and life in general. These measures include the enhancement of education and training programmes, provision of an environment conducive to the development of ICT, provision of incentives for computerization and automation, and creation of venture capital funds. Currently, Malaysia is in full gear to steer the economy towards a knowledge-based one.

A national broadband master plan to enable the country to have a 50% penetration rate for broadband services by the year 2007 has also been implemented. The government sees enhanced networking as vital for the e-learning initiative undertaken by the government to accelerate the growth of education in the country and in ensuring that the country makes the transition towards becoming a knowledge-based society.

The Ministry has formulated three main policies for ICT in education. Whereby, the second policy emphasizes on the role and function of ICT in education as a teaching and learning tool, as part of a subject, and as a subject by itself. Apart from using radio and television as a teaching and learning tool, this policy stresses the use of the computer for accessing information, communication, and as a productivity tool. ICT as part of a subject refers to the use of software (e.g. AutoCAD and SCAD) in subjects such as "Invention" and "Engineering Drawing." ICT as a subject refers to the introduction of subjects such as "Information Technology" and "Computerization".

Despite of the efforts carried out by the Malaysian government on ICT, according to Schank (2007), modern technology has had very little effect on educators' conceptions of teaching and learning. Besides that, institution authorities have spent millions of ringgit in investment to equip their centres with educational technologies such as computer lab, LCD projector, networking or other computer peripherals like printers and modems to assist teaching and preparations of teaching materials. Moreover, some have engaged professionals to give computer courses to their academic staff in preparation to step up as world-class university. As indicated and found in a few studies cited below, this survey would like to look into the Malaysian Higher Institution scenario of ICT utilization among their educators.

Through Internet and accredited technology journals, it is widely discussed and known that higher institutions and schools in America and other developed countries had been integrating the technology into their classrooms (Neo et. al, 2001, Al-Seghayer, 2001 and Nikolova, 2002). The benefits of it are undoubtedly very significant. Students are said to be more eager and highly motivated (Samad, 1997) because they can access their learning anywhere and at anytime provided they have a computer. A study conducted by Chris Rother (2003) a vice president of Education of CDW.G found that 86% of their respondents who are students said in class computers have improved their academic performance and 74% said it has increased their attention in class. In fact, 65% of the teacher respondents who responded to the survey said that computers can be more effective than teachers in conveying certain types of information to the students (Rother, 2003). However, the effectiveness of ICT usage still depends on the teachers and the students in which both parties must be interested and willing to engage with computers (Jones, 1999) and how teachers integrate computer activity in a meaningful learning activity (Demetriadis, 2003). Computers alone might not be sufficient as it needs the integration of the technology product and activities (Samad, 1997). A study conducted by Davis et al (1997) found the quality of the learning can be significantly enhanced when ICT is integrated with teaching.

In reality, most instructors have some familiarity with computers and are able to use a variety of computer softwares as found in one study done by the National Education Association. It was found out that 94% of all respondents in the survey are able to search the Internet. However, they do not know how to fuse computer skills into classroom instruction. As a study conducted by Cuban (1999) reported, out of every 10 teachers in U.S., fewer than two seriously are users of computers and other information technologies in their classrooms (several times a week); three to four are occasional users (about once a month); and the rest (4-5 teachers out of every 10) never use the machines at all. As another findings from a survey on Survey of ICT utilization in Philippines Public High School' stated that 92% of the respondents who are teachers of the public school said that there is a need for more information given to them on how to use ICT to support the curriculum and 96% of the respondents need to develop skills to hands on activity to share with their students (Tinio, 2002). A study conducted by Ahmad (2007), in Open University Malaysia (OUM) in 2006, found out that '….the more

senior learners prefer mostly face-to-face interactions and are not too comfortable with online courseware and interactions.' However this finding contradicts with a research findings conducted by Bee Theng, Lau and Chia Hua, Sim (2008), whereby the age has a negative relationship with the extent of ICT use among teachers. Senior teachers were found to be highly positive towards ICT use in their teaching and professional work, and had translated this into a greater use of ICT in schools. The teachers' computer competency have an overall mean of 3.35 (SD=0.71), indicates that teachers generally feel competent in utilizing ICT tools in school. Another study on ICT implementation in Malaysia, conducted by Shamsul et.al (2006) focused only on the implementation of ICT vision, plan policies and strategies.

Technology can play various instructional roles - and it is the responsibility of the instructors to decide how to best use technology to support student learning. Having a complete infrastructure of the ICT will go meaningless if it is not utilized to the fullest capacity. Meanwhile, Schwach (2004) and Demetriadis et. al (2003) argue that the effective use of technology in classroom is not only limited to the teachers' perceptions on how to use technology is class but also through professional development for teachers. Their study indicates that training is needed in order for teachers to be able to integrate computer in their classroom practice.

The general objective of the study is to explore on the ICT utilizations among the IPTA and IPTS educators of northern Malaysia. The specific objectives of this study are three folds: a) to investigate the extent of ICT resources provided by universities authorities, b) focus on types and extent of ICT usage in daily activities, c) to explore the ICT proficiencies level and d) to investigate the level of ICT integration in teaching activities.

2. Method

A survey instrument was designed to gather information on the computer usage adapted from two previous studies done by Victoria L. Tinio in 'Survey of Information and Communication Technology Utilization in Philippine Public High Schools' in 2002 and 'Faculty Responses on the Status of Technology at Southern Mississippi University' in March 2001. The questionnaire was edited and rephrased to suit our research objectives and the infrastructure of Malaysia Public and Private Higher Learning Institutions.

All together there were 90 questions to be answered. The types of questions used in the questionnaire vary from close-ended, scales to matrix questions. A reliability test was also carried out to determine the internal consistency between items used in the questionnaire namely Cronbach's Alpha and all of the questions asked had values of 0.7 and above. It means that they were relevant and significant to our objectives.

Frequency and duration are the most common scales used to measure usage. The questionnaire requires the respondents to provide the estimated time spent daily on computer to perform certain activities. Based on pilot study revealed that the duration of individual session on the computer were highly variable. Therefore, duration was used as the operational definition of usage categories. Meanwhile, frequency was used to give a measure of the specific types of activities for which the respondents used the computer. These activities include integration of computer technology in teaching, instructional activities, communication, organizational activities, creative, expressive, evaluative and informative.

The instrument also measure proficiency level related to computer technologies and integrating information technology in teaching activities. The proficiency level was categorized into five, namely; unfamiliar, beginner, average, advance and expert. Again, five categories have been identified to determine the process level in integrating computer technology namely, awareness, learning, familiarity, adaptation and creative application.

A census survey was conducted on the educators of Faculty A in IPTA and IPTS in the Northern Region of Malaysia which includes the states of Kedah, Penang, and Perak. There are four public higher learning institutions (IPTA – Institusi Pengajian Tinggi Awam) and seven private higher learning institutions (IPTS – Institusi Pengajian Tinggi Swasta) that were chosen in this study. From the 250 number of respondents of IPTA, 76 responses have been received. This is about 30.4% of the population. Whereas, a total of 280 questionnaires were distributed to 7 IPTS's and only 108 were returned. This is about 51.9% of the population.

3. Results and discussion

The findings will be presented in the order of the level of ICT resources, types and extent of ICT usage, level of ICT proficiencies and level of ICT integration in teaching activities.

The professional view on ICT usage in classroom among educators relies heavily on the extensiveness of computer resources availability at their premises. The educators also fairly strongly agree that ICT is a valuable instructional tool and by utilizing it in their preparation for teaching materials and in class teaching will enhance their professional development. They also agree that utilizing ICT in the curriculum will boost their confidence as a competent educator. Furthermore, they also believe that it will promote their development of communication skills in writing and presentation.

3.1 Level of ICT resources

The survey also uncovered the fact that the level of ICT resources is still inadequate for academic use for educators. With a mean score of 3.79 for IPTA and 3.9 for IPTS, it shows that the educators' access to those resources is insufficient

for their educational purposes. This is due to the facts that, majority of the institutions do not have enough ICT resources provided. While according to Pedro (2005), heavy investment in ICT must be taken seriously by university authority in order to improve the teaching quality. He further stated two reasons for such investment. The first reason is university education has a responsibility to ensure the future graduates are well versed in the use of ICT, since in a knowledge economy; such technologies are very important tools of every day life when a student enters the work market. The second reason is that ICT may contribute to more and better learning to improve the effectiveness of university education.

Our finding also revealed that other ICT resources which are considered as important received a low mean score such as LAN (local area network) with 3.49 for IPTA and 3.39 for IPTS. The lowest score is WiFi, receiving a mean score of 2.54 for IPTA and 2.03 for IPTS. (Table 1)

However, as Unwin (2007) observes, "… it is not the availability of the technology which is important, but how it is used" that matters. In order to encourage them integrate the technology into the curriculum, enough resources should be made available to them besides providing courses and workshops to assist them master the related software according to their discipline.

3.2 Level of ICT usage in daily activities

Eight aspects of computing purposes were identified in the initial study as being regularly used with the educators: informative, communicative and expressive, integrating computer technology, evaluative instructional, organizational and creative purpose. This study compares the usage frequencies between the IPTA and IPTS educators.

Most of the respondents from IPTA and IPTS are dedicating their daily activities for informative, communicative and expressive purposes (i.e., nearly every day). The informative purposes might include activities such as searching for information over the Internet and CD-ROM; yield the highest frequency for both IPTA and IPTS with percentage response of 60.8% and 51.9% respectively. Then followed by communicative purposes such as sending/receiving e-mail, ICQ, computer conferencing and using LCD projector and expressive purposes which include activities such as word processing (typing, editing, layout), and slide presentation. (Table 2)

The result is consistent with a study done by Chong et al (2005) that showed most of the educators use computer on a regular basis for common computer packages such as word processing, spreadsheet, and for internet services such as search engine. This is supported by a study by Yasmin, Wan Suriyani and Sidi Merican (2008) who found educators in UniKL commonly used computer slides presentation and reading materials from web site.

Surprisingly, only fewer number of respondents claimed of using higher level skills activities such as evaluative (e.g. assignments, portfolio, testing), instructional (e.g. drill practice, tutorials, remediation), organizational (e.g. database, spreadsheets, record keeping, lesson plans) and creative (e.g. Desktop publishing, digital video, digital camera, scanners, and graphics) as these activities require specialized knowledge and training in order to use them. However, these activities are vital to educators who really want to incorporate the technology skills with the understanding of the teaching and learning. This finding supported a study which was conducted by Asrun et al (2003) and Castillo (2005) that showed majority of the educators do not use new opportunities that are available in ICT such as graphic application, multimedia and some authoring applications in teaching activities.

3.3 Level of ICT proficiencies

Considerable differences were found in the levels of proficiency between the two institutions. Even though the IPTA outnumbered IPTS in terms of ICT resources, however, IPTS outperformed IPTA in terms of ICT level of proficiency. In the IPTA, 48.6% of lecturers are at average level of ICT proficiency, as compared to 44.8% of IPTS educators are in the advanced level. Surprisingly, they are 2.8% of IPTA and 4.8% of IPTS were in the beginner level despite of their seniority in service. (Table 3)

The study also found that, though the educators claim they are expert / advanced in the level of proficiency, however the ICT resources that can support them into becoming creative applicants of teaching and learning process are insufficient.

3.4 Level of ICT integration in teaching activities

The study also attempted to determine the educators' process level in integrating computer technology in their teaching activities. (Table 4) The highest response for both IPTA and IPTS educators fall under the adaptation category. Here the respondents think about the computers as an instructional tool to help them perform appropriate tasks. They are no longer concern about it as a technology. However, educators' who are in this category are able to use many different computer applications to aid them in delivering the required knowledge to their students. Quite a number of the educators from IPTA (around 31.9%), as compared to 24.04% of the educators from IPTS indicated they are in the category of familiarity, which is beginning to understand the process of using technology and can think of specific task in which the technology might be used. Whereas, more IPTS educators were (29.81%) under the creative application category, as compared to IPTA educators (22.2%). In this category, they claim that they can apply their knowledge about the technology in the classroom and able to use them as an instructional aid and able to integrate computers into their curriculum. Among the

IPTS respondents, none reported as having awareness level of proficiency, but surprisingly, there are 1.4% of IPTA educators who were just aware that the technology exist but have not used it or perhaps avoiding the technology. This finding is consistent with research done by Asrun (2003) which showed that even though educators have positive attitudes towards ICT usage, however they are not convinced that the use of ICT in teaching will lead to better student outcomes.

Unfortunately, only 22.2% of IPTA respondents versus 29.8% of IPTS respondents are actually applying their technology expertise into their learning/teaching process. A majority of them just 'believe' they can use technology in doing their job, but not integrating them yet. We may suggest that this is happening due to the attitude that educators still hold. These educators may feel "very unsure about the effective use of technology" or believe that "computer activities are just a waste of time" (Fryer, 2004).

Another possible reason for such reluctance as Zhang (2007) "…argues that the pedagogical cultures of the East and West are in conflict, causing certain reluctance among Asian nations to adopt ICTs because they are so closely connected with the theory of constructivis."

4. Conclusion

From the study conducted, the following conclusions can be derived: a) ICT resources at both institutions are still unsatisfactory to facilitate the teaching and learning activities. b) In terms of ICT usage, most of the educators from both institutions dedicated / spent most of their daily activities for informative, communicative and expressive activities. c) Level of ICT proficiencies among the IPTS educators are higher than the IPTA whereby a majority of IPTS educators claimed to be in the advanced level unlike a majority of IPTA educators who were just in the average level. d) In terms of the level of ICT integration process, most educators from both institutions are under the adaptive category. But a few of the IPTA educators were reported as 'never use' or 'rather avoid' using ICT in their teaching activities.

References

Ahmad Hashem (2008). ICT Initiatives in Higher Education: The Open University Malaysia (OUM) Perspectives. In ICT Conference and Exhibition, 10-12 March 2008. Kuala Lumpur, Malaysia. Retrieved 22 July 2009 from http://www.moe.gov.my/43seameocc/download/MSIA-OUM.pdf

Allan H.K. Yuen, Nancy Law and K.C. Wong (2003).ICT Implementation And School Leadership: Case Studies Of ICT Integration In Teaching And Learning," *Journal of Educational Administration,* 41(2): 2003, pp. 161. Retrieved 6 December 2007 from http://www.emeralinsight.com/0957-8234.htm

Al-Segghayer, K. (2001). The Effect of Multimedia Annotation Modes on L2 Vocabulary Acquisition: A Comparative Study. *Journal of Language Learning & technology*, 5(1): pp.202-232.

Asrun, M., Dal, M. and Samuel, C. (2003). How do Teachers Use Information and Communication Technology in Iceland high Schools in 2002? International Conference on Computer Science and Technologies – CompSysTech'2003. Sofia, Bulgaria: 19-20 June 2003. Retrieved 2 March 2008 from http://ecet.eces.ru.acad.bg/cst/Docs/proceedings/S4/IV-4.pdf

Bee Theng, Lau and Chia Hua, Sim. (2008). Exploring The Extent of ICT Adoption Among Secondary School Teachers in Malaysia. International Journal of Computing and ICT research. Vol. 2, No. 2, pp.19-36. Retrieved 22 July 2009 from http://www.ijcir.org/volume2-number2/article 3.pdf.

Castillo, N.(2005). The Level of ICT Use and Expertise by Teachers in Chilean Secondary Schools. Retrieved 15 January 2008 from http://www.tise.cl/archivos/tise2006/23.pdf

Chan, Foong-Mae., (2002). ICT in Malaysian Schools: Policy & Strategies. In Seminar / Workshop on The Promotion Of ICT Education to Narrow the Digital Divide, 15-22 October 2002. Tokyo, Japan.

Chong, C.K., Sharaf Horani and Jacob Daniel, (2005). A study on the use of ICT in Mathematics Teaching. *Malaysian Online Journal of Instructional Technology(MOJIT)*, 2(3): pp 43-51.

Cuban, L., (1999, 4 August). The Technology Puzzle. *Education Week,* 47, 68. Retrieve 22 November 2003 from http://www.edweek.org/ew/vol-18/43cuban.h18.

Demetriadis, S., et.al., (2003). Culture In Negotiation: Teachers' Acceptance/Resistance Attitudes Considering The Infusion Of Technology Into Schools. *Computers & Education*, 41, pp.19-37 in the article of Nadzrah Abu Bakar & Peter Mickan,(2004). Students' Experience In Computer-Based English Language Classroom, *Proceedings Of The 2003 ASIACALL International Conference On IT And Language Education*, pp.1-7. Korea.

Jones, J., (1999). Language Learning, Technology and Development: The Essential Interaction Between Teachers And Students. *The Fourth International Conference on Language and Development* in the article of Nadzrah & Peter, (2004). Students' Experience In Computer-Based English Language Classroom, *Proceedings Of The 2003 ASIACALL International Conference On IT And Language Education*, pp: 1-7. Korea.

National Education Association (2000). Retrieved October 1 2006 from http://www.unionfacts.com/unions

Neo, M. and K.T.K. Neo. (2001). Innovative Teaching: Using multimedia in a problem-based learning environment. Educational Technology & Society 4(4)

Nikolova, O.R., (2002). Effect of Students' Participation in Authoring of Multimedia Materials on Students Acquisition of Vocabulary. Language Learning & Technology.6(1): pp.100-122

Pedro, F.(2005). Higher Education in Europe. Comparing Traditional and ICT-Enriched University Teaching Methods: Evidence from Two Empirical Studies. 30(3-4).

Rother, C.,(2003). Technology's Value in Education. Retrieved September 15 2004 from http://www.thejournal.com/articles/

Samad, R.S.A., (1997). Teknologi Mencambah Minat Pembelajaran. Dewan Budaya. 19(2): pp. 48-49

Schank, R. (2007). Teching In The New Era. In C. Crawford, R. Carlsen, K. McFerrin, J. Price, & R. Weber (Eds.). Proceedings of Society for Information Technology and Teacher Education International Conference 2007 (Keynote). Chesapeake, VA:AACE.

Schwach, E. (2004). How Do Professional Development Opportunities Assist Teachers In Using Technology In Their Classrooms? 2001, http://users.rcn/com/tinshee/paper.htm in the article of Nadzrah Abu Bakar & Peter Mickan, Students' Experience In Computer-Based English Language Classroom, Proceedings of the 2003 ASIACALL International Conference on IT and Language Education, Korea, pp. 1-7.

Shamsul, A. M., Rose, A. A., Azizah, A. R., (2006). Rubric For Assessing ICT Vision, Plan, Policies and Standards in Malaysian Higher Education. International Journal of Education and Development using Information and Communication Technology. Vol.3, Issue.2 (2007). pp. 30-56

Tan, C. (2002). Towards A Smart Nation. Jurutera, 2002(4): pp. 6-12

Tinio. Victoria L. (2002). Survey of Information and Communication Technology Utilization in Philippine Public High Schools. Retrieved February 15, 2004 from www.digitalphilippines.org/research_fullarticle

Unwin, A. (2007). The Professionalism of The Higher Education Teacher: What's ICT Got To Do With It? Teaching in Higher Education. 12(3). Pp. 295-308. In Wang, T. Rethinking Teaching With Information and Communication Technologies (ICTs) In Architectural Education. Teacher and Teacher Education (2009), doi: 10.1016/j.tate.2009.04.007

Table 1. Level of ICT resources

Level of ICT resources	**IPTA mean**	**IPTS mean**
Instructor access computer	3.79	3.9
Computer availability	3.53	3.16
Local area network	3.49	3.39
WiFi	2.54	2.03

Table 2. Level of ICT usage in daily activities

Level of ICT usage in daily activities	**IPTA % of response**	**IPTS % of response**
Informative	60.8	51.9
Communicative	50.7	41
Expressive	46.6	41.9
Evaluative activities	32	24
Instructional	25.3	22.9
Organizational	22.4	28.6
Creative	6.7	11.5

Table 3. Level of ICT proficiencies

Level of ICT proficiencies	IPTA	IPTS
	% of response	% of response
Beginner	2.8	4.8
Average	48.6	43.8
Advanced	34.7	44.8
Expert	13.9	6.7

Table 4. Level of ICT integration in teaching activities

Level of ICT integration	IPTA	IPTS
	% of response	% of response
Awareness	1.4	0
Learning	1.4	8.65
Familiarity	31.9	24.04
Adaptation	43.1	37.5
Creative Application	22.2	29.81

Computer and Information Science
www.ccsenet.org/journal.html
Vol. 2, No. 4
November 2009

A NN Image Classification Method Driven by the Mixed Fitness Function

Shan Gai, Peng Liu, Jiafeng Liu & Xianglong Tang
School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
E-mail: gaishan@hit.edu.cn

The research is financed by the National Natural Science Fund of China (No.60702032) and the Natural Science Innovation Fund of Harbin Institute of Technology (No.HIT.NSRIF.2008.63). (Sponsoring information)

Abstract

The mixed fitness function of the error sum squares linear transformation is proposed in the article, and this function can improve the evaluation method of the individual fitness, and combining with NN, this method can be used in the high-speed paper money image analysis system. Aiming at many characters such as the high comparability of paper money images of different denominations, small class distance and large in-class discreteness induced by the using abrasion, this method first codes the weight values and threshold values of NN with real values, and transforms the problem from the representation type to the genotype, and performs many genetic operations such as selecting, crossing and variation, and takes the weight value and threshold value trained by the genetic algorithm according to the individual fitness value of the mixed fitness function as the initial weight value and initial threshold value of NN in the next stage, and trains these values by NN to establish the sorter. This method was tested in the embedded system with resource restriction (TI TMS320C6713 DSP), and 20000 RMB images were acquired as the samples, and 12000 images of them were tested, and the test result indicated that the method combining improved genetic algorithm with NN obviously enhanced the recognition rate.

Keywords: Mixed fitness function, Genetic algorithm, NN, Paper money image classification

1. Introduction

Financial institutions would settle large numbers of paper money every day, which requires that the paper money sorting system possesses quick processing speed and high identification reliability. The paper currency sorter is an automatic settlement took, and it is used to sort the denomination, face and deformity class of paper money. The designs of paper money with different denominations are similar, and the pollution and depreciation will make the discreteness of the money samples with same denomination and face very large, and the running instability of the high-speed equipment will change the geometric shape of paper money, so the geometric size of the paper money can not be a reliable classification reference. The image analysis technology is the core technology in the paper money sorting system, and to use the image to identify the denomination of paper money in the financial tools is a new identification measure in recent years.

The design of the sorter is an important stage in the identification of paper money, and the minimum distance sorter has been applied in the paper money identification system, and it could compute the distances between unknown samples and each training sample vector in the models, and select the minimum distance among them to make a decision, but its efficiency is low. So Takeda and Omatu (Fumiaki Takeda, 1995, P.73-77) applied NN in the sorter design of paper money identification in 1995, and acquired better effect. Ahmadi (Ahmadi A, 2003, P.2550-2554) et al applied the learning vector quantification (LVQ) (Ahmadi, 2004, P.1313-1316) into the sorter design. As the sorter of the paper money identification, the NN algorithm possesses many advantages such as parallel processing, generalization, self-organization and exact optimization. Because the BP algorithm adopts the search solution algorithm solution descending along the grads, and the learning speed decides the weight value change in cycled training, and large learning speed may induce the instability of the system, the error squares will fluctuate, and the local minimization and slow convergence speed will be induced in the network.

The basic idea of the genetic algorithm (GA) is to adopt certain coding mode to map the solution space to the coding space, and each code corresponds with one chromosome or individual. The random method is used to confirm one

initial group of individual, which is called the species group. In the species group, the individual is selected according to the fitness or certain competition mechanism, and the genetic operators such as selecting, crossing and variation are used to generate the next generation, and in this way, the evolvement continues until the condition fulfilling the expectation ends. Barrios et al improved the coding method (Barrios, 2000, P.844-847), and Miller improved the operating operator of the GA (Miller, 1993, P.1340-1351), and the standard GA is improved from the design of the fitness function in the article, and a mixed fitness function combining error squares and linear transformation is proposed. The mixed fitness function can solve the problem of the individual with super-large fitness value in the species group when singly adopting the error squares, and effectively extend the range of the fitness value. The GA has strong macro searching ability, and it can find the global optimal solution by the big probability, and it has high robustness. But GA is deficient for the local searching ability, and its execution efficiency is low and the convergence will occur too early.

According to the disadvantages and advantages of BP (back propagation) and GA, a mixed GA-BP algorithm has been successfully applied in many domains such as earthquake prediction (Qiuwen, 2008, P.128-131) and sound identification and acquired better effect (Min-Lun, 2006, P.527-530). The GA-BP is applied in the paper money identification in the article, and because the images of paper money with different denominations are very similar, the class distance is small and the in-class discreteness is large, this article first codes the weight value and threshold of BP network with real values, and then uses GA to train the weight value and threshold values of the network (Jiansheng, 2005, P.288-291 & Chien-Yu, 2008, P.1459-1465), and uses the BP algorithm (Fariborz, 2008, P.389-404 & Ming, 2008, P.115-119 & Qiang, 2005, P.357-360) to train the weight value and the threshold value optimized by GA, and finally predict the unknown samples. The test result shows that the mixed algorithm combining improved genetic algorithm and BP network in the article has higher identification rate and reliability.

2. Basic principles of BP algorithm and genetic algorithm

NN is one method to solve the nonlinear problem, and it has strong function approaching complex nonlinear function and strong fault-tolerant ability, so it can establish the NN model by existing sample information, and with the continual accumulation of sample information, it can perform self-study based on new sample information, and form more perfect and exact evaluation system. BP network is a kind of multi-layer feed forward NN, and if the input and output relations are gave continually, the interior will certainly form the internal structure with this relation in the learning process of BP network. Each neutron in BP network has several outputs, and it connects with many other neutrons, and each neutron corresponds with several connection accesses and each connection access corresponds with one connection weight value coefficient. Each node in the network has one status variable x_i and one threshold variable θ_j. The connection weight coefficient from the node i to the node j is w_{ji}. For each node, one transformation function $f(x)$ is defined, and generally, $f(x) = 1/1+\exp(-x)$. When the input vector and the objective output vector are confirmed, through initial connection weight coefficient and the threshold value, the network performs the nonlinear reasoning according to the transformation function, and according to the error between the obtained actual output vector and objective output vector, the connection weight efficient and threshold value are adjusted. Through the repeating training to above process, when the error between the actual output vector and the objective output vector achieves the error pre-established, the training process ends. So the connection weight coefficient and threshold value among adjusted nodes can be used to predict unknown samples. The computation process of BP network is divided into the input mode forward propagating and the output error reverse propagating. And the process of the forward-propagating can be described as follows.

$$s_j = \sum_{i=1}^{n} w_{ji} \cdot x_i - \theta_j \qquad (j = 1, 2, \cdots, p) \tag{1}$$

$$b_j = f(s_j) = \frac{1}{1 + \exp\left[-\sum_{i=1}^{n} w_{ji} \cdot x_i + \theta_j\right]} \tag{2}$$

Where, s_j denote the activation value of various neutrons and the activation function adopts above S-type function, b_j denotes the output of the j'th unit in the hidden layer, and p is the amount of neutron in the hidden layer in the network. In the same way, the activation value and the output value can be solved.

$$s_k = \sum_{j=1}^{p} v_{kj} \cdot b_j - \theta_k \qquad (k = 1, 2, \cdots, q) \tag{3}$$

$$y_k = \frac{1}{1+\exp\left[-\sum_{j=1}^{p} v_{kj} \cdot b_j + \theta_k\right]} \tag{4}$$

The mode forward propagating is used to obtain the actual output value of the network, and when the error between the actual output value and the expectation output value is big, the connection weight value and the threshold value in the network should be modified. The error reverse propagating process of the BP network can be described as follows.

$$d_k = (o_k - y_k) y_k (1 - y_k) \qquad (k = 1, 2, \cdots, q) \tag{5}$$

$$e_j = \left[\sum_{k=1}^{q} v_{kj} d_k\right] b_j (1 - b_j) \qquad (j = 1, 2, \cdots, p) \tag{6}$$

Where, d_k denotes the correction error of the output layer, o_k denotes the expectation output, e_j denotes the correction errors of various units in the hidden layer, and the formula (5) and the formula (6) can adjust the connection weight values and the threshold values layer by layer from the output layer to the hidden layer, and from the hidden layer to the input layer.

$$\Delta v_{kj} = \alpha \cdot d_k \cdot b_j \tag{7}$$

$$\Delta \theta_k = \alpha \cdot d_k \tag{8}$$

$$\Delta w_{ji} = \beta \cdot e_j \cdot x_j \tag{9}$$

$$\Delta \theta_j = \beta \cdot e_j \tag{10}$$

Where, Δv_{kj} and $\Delta \theta_k$ respective denote the connection weight value corrections and the threshold corrections from the output layer to the hidden layer, $\alpha(\alpha > 0)$ denotes the learning coefficient, Δw_{ji} and $\Delta \theta_j$ respectively denotes the connection value corrections and the threshold value corrections from the hidden layer to the input layer, and $\beta(0 < \beta < 1)$ denotes the learning coefficient.

GA is a kind of probability searching algorithm, and it utilizes certain coding technology to act on the number cluster which is called chromosome, and its basic idea is to simulate the individual evolvement process composed by these clusters. Its essential is a kind of high-effectively, parallel and global searching algorithm, and it can automatically acquire and accumulate knowledge about searching space in the searching process, and self-adaptively control the searching process to seek the optimal solution. The GA uses the principle of the survival of the fittest, and gradually generates an approximately optimal project in the potential solution group. GA first performs coding, i.e. realizes the mapping of problem from the representation type to the genotype, and then calculates the fitness function, and its value reflects the situation of the individual, and finally the genetic operators such as selecting, crossing and variation are calculated, and the approximately optimal solution of the problem can be sought.

Because the BP algorithm has the self-organization and self-study abilities, it can directly accept data to perform the study, and self-adaptively find the rule in the sample data, and it has good extension ability to introduce new money types in the paper money identification system. So the BP algorithm is very fit to be used in the processing of paper money image. But the BP algorithm is easily to get in local optimization, slow convergence speed, and uncertain initial weight value and threshold value of the network. GA is a global optimal searching technology, and it can effectively compensate the disadvantages of the BP algorithm. In the paper money identification, GA is used to seek the optimal initial weight value and threshold value in the network and according to the character of similar paper money images, the BP algorithm is used to train the weight value and the threshold value.

3. NN based on improved genetic algorithm

3.1 Improved genetic algorithm

The fitness value in the GA is used to measure the degree that various individuals achieve or approach the optimal solution in the optimization computation. The individual with high fitness has larger probability to be inherited to next generation, and the individual with low fitness has relative lower probability to be inherited to next generation. The function to measure the individual fitness is the fitness function, and it is the drive of the GA evolvement, and the unique reference to perform natural selection. The extensive fitness function in GA is the error squares. But if in the initial group, certain special individual with excessive fitness exists, this function can not prevent this individual to govern the group, and it will mislead the optimization development direction of the group, and make the algorithm to be convergent in the local optimal solution, and when the GA is gradually convergent, the individual fitness values in the group will be close, and it will be difficult to perform the optimization, and the optimal solution will be easily to sway

near the optimal solution. Based on above reasons, a linear transformation of the fitness function is introduced in the article, and its intention is to properly amplify the value of the fitness, and increase the selecting ability of GA. The concrete linear transformation formula is

$$f' = \left(f - f_{\min}\right) \Big/ \left(f_{\max} - f_{\min}\right) \tag{11}$$

Where, f is the original fitness value, $f_{\min}$ is the lower limit of the fitness function value, $f_{\max}$ is the upper limit of the fitness function value, and f' is the fitness value after transformation. From the formula (11), if the difference between $f_{\max}$ and $f_{\min}$ is big, the fitness value after transformation will be correspondingly reduced, which can effectively avoid the problem misleading the group optimization direction because of the existence of the individual with super-large fitness. But the linear transformation has not same effect to describe the difference among similar individuals than the error square, so a mixed fitness function such as the formula (12) is proposed as follows.

$$f'' = a \cdot \frac{1}{E} + (1-a) \cdot f' \tag{12}$$

Where, E is the sum of error square, and a $(0 \le a \le 1)$ is the harmonic coefficient. The formula (12) fully considers the knowledge of the concrete problem field of the paper money, and adds the information of the change rate of the fitness function value into the fitness function, which can effectively overcome the limitation that the chromosome selected in the standard GA may be not good chromosome, and avoid the phenomena of earliness, and possess higher convergence speed. The harmonic parameter a in the formula (12) can be finally confirmed by the test mode.

3.2 Optimizing NN by genetic algorithm

The advantage of BP algorithm is that it is easy to be implemented and the optimization is exact. But it has two disadvantages. First, the BP algorithm is easy to get in the local minimization, because the error curve generally has several extreme points. Second, the convergence speed of the BP algorithm is slow, and when the grads descending method is adopted, the step length is difficult to be confirmed, and if the step length is too large, the required precision can not be achieved, and even the result will be dispersed, and if the step length is too small, the iterative approach will increase and the convergence speed will decrease. The advantage of GA is that it can not easily get in local optimization in the searching process, and even if the defined fitness function is not continual and regular, or has noise, it can find the global optimal solution by the large probability, and possess strong robustness. At the same time, the GA has many disadvantages such as low efficiency, too early convergence and weak local searching ability.

Therefore, a mixed algorithm (GA-BP) combining above two algorithms is formed in the article to optimize the NN. Its idea is to first train the weight value and threshold value of BP network by GA which replaces the method randomly evaluating the connection weight value and threshold value by the standard BP network, and can effectively reduce the searching range, and then train the weight value and threshold value optimized by GA by the BP network, and finally utilize the generalized ability of the network to predict the input unknown samples. The approaches of the GA-BP training algorithm can be described as follows. Initialize the species group and crossing probability Pc, the variation probability Pm, the weight values w_{ji} and v_{kj}, the threshold values θ_j and θ_k, and perform the coding by the real numbers, and the coding length is seen in the formula (13). In the formula (13), S_{in} denotes the amount of neutron in the input layer, S_{out} is the amount of the neutron in the output layer, S_i is the amount of the neutron in the hidden layer i, and p is the amount of the hidden layer.

$$S = S_{in} + \sum_{i=1}^{p} S_i + S_{out} \tag{13}$$

The formula (13) can realize the transformation of the paper money image from the representation type to the genotype. Then compute the fitness functions of the individuals, and rank them, and select the individuals in the initial species group according to the probability value of the formula (14).

$$p_j = f_j \Big/ \sum_{j=1}^{N} f_j \tag{14}$$

Where, f_j is the fitness value of the individual j, and it can be measured by the formula (15).

$$f_j = a \cdot \frac{1}{\sum_p \sum_k \left(y_k - o_k\right)^2} + (1-a) \cdot \frac{f_j - \max\limits_{1 \le j \le N}(f_j)}{\max\limits_{1 \le j \le N}(f_j) - \min\limits_{1 \le j \le N}(f_j)} \tag{15}$$

Where, $j=1,2,\cdots,N$ is the amount of chromosome, p is the amount of sample, k is the amount of neutron in the input layer, y_k is actual output of the network, and o_k is the expected output of the network. Use the crossing probability Pc to perform the crossing operation to the individuals c_j and c_{j+1} to generate new individuals $c_j^{'}$ and $c_{j+1}^{'}$, and the individuals which are not be crossed will be copied directly. Utilize the variation probability P_m to generate new individual $c_i^{'}$, and insert the new individual to the species group, and compute the new fitness function value. If the new individual fulfills the conditions, the optimization ends, or else, continual perform the genetic operator operation to the group. Finally perform the decoding operation to the individuals in the final group, and obtain the optimized connection weight value and threshold values of NN. Aiming at the characters of the paper money image, the GA-BP algorithm can be adopted to train the samples in the network under the optimal initial weight value and threshold value, enhance the performance of the network, quicken the convergence speed, and avoid getting into the local optimization.

4. Test result and conclusion analysis

4.1 Establishment of test database

To validate the efficiency of the method in the article, 20000 paper money images were collected in the multi-function money detection instrument designed, and the light-source sensor with 200dpi resolution, and the samples include five types of paper money of 2005 edition RMB. Each money type has four faces and there are 20 classes. Each class has 1000 samples, and 400 of them are used for training, and other 600 samples are used for test.

4.2 Preprocessing of paper money image

The image should be positioned before abstracting the character of the image, i.e. finding the position of the paper money image. In the article, test many dispersed points on the borders of the papery money first, and then adopt the least square method to fit the border line of the paper money image for the border sequence points (seen in Figure 2). Because the paper money image collection is to scan the image in the paper money movement, so the geometric distortion will generally occur to some extent, and this distortion comes from two aspects, and one aspect is induced by the slope of the paper money, and the other aspect is induced by the transverse movement in the scanning process. The slope correction of the paper money image is seen in Figure 3.

4.3 Character abstraction of paper money image

The meshing character is adopted as the identification character, and the size of the collected paper money image is 270×150 pixel. Through analyzing of the paper money images with different types, the sensitive region with predominant contribution to the identification can be confirmed. These regions are divided into small panes of 16×6 in the article, so each paper money will form 96-dimensional eigenvector, and the eigenvector of each dimension is the sum of pixel grey value of the corresponding pane, and standardize the output to obtain the eigenvector.

4.4 Analysis of test result

The intention of the test 1 is to confirm the harmonic coefficient a of the mixed fitness function, and 12000 samples were tested in the article, and the optimal value is confirmed through setting up corresponding identification rate of the different harmonic coefficient. From Table 1, when $a=0.6$, the identification rate is highest, and when the value of a increases or decreases, the identification rates all will decrease. Figure 5 are the corresponding mixed fitness function error curve and the fitness function value curve of different harmonic coefficients, and when $a=0.6$, the convergence speed of the convergence speed is quick, and the fitness function value curve can use less iterative times to achieve the stable state.

In the article, the GA fitness function adopts the error square as the standard GA, which is denoted by SGA, and the fitness function adopts the mixed function as the improved GA, which is denoted by IMGA. The intention of the test 2 is to compare the performances of BP network in Omatu's article (Omatu, 2007, P.413-417), the combined network of SGA and BP, and the combined network of IMGA and BP in the paper money identification. The fitness error function of IMGA in Figure 6 has quicker convergence speed than the fitness error function of SGA, and the iterative times that the fitness value goes to stable is less than the iterative times of SGA. The data in Table 2 indicates that the identification rate using the combination of IMGA and BP network is higher than the identification rate singly using BP network or the combination of SGA and BP network.

5. Conclusions

BP NN has the problem of local minimization and GA has good global searching ability, so an improved GA is proposed in the article, and a mixed GA-BP algorithm combining improved GA and BP NN is applied into the paper money identification. GA-BP mixed method could optimize the find the optimal point in the solution space, and search the BP network according the negative grads direction, which can avoid that the BP algorithm gets into the problems

such as local minimization and slow convergence speed, and overcome the disadvantages of GA that the searching time is too long and the searching speed is slow in the optimization process. The test result indicates that the algorithm in the article has higher reliability and robustness in the paper money identification.

References

Ahmadi, A, Omatu, S, & Kosaka, T. (2003). A study on evaluating and improving the reliability of bank note Euro-sorters. *SICE 2003 Annual Conference.* P.2550-2554.

Ahmadi, A, Omatu, S & Kosaka, T. (2004). Improvement of the reliability of bank note sorter machines. *Proceedings of the IEEE International Conference on Neural Networks.* P.1313-1316.

Barrios, D, Manrique, D, Porras, J & Rios, J. (2000). Optimum binary codification for genetic design of artificial neural networks. *Proceedings of the Fourth International Conference on Knowledge-Based Intelligent Engineering System and Allied Technologies.* P.844-847.

Chien-Yu Huang, Long-Hui Chen, Yueh-Li Chen & Fengming M. Chang. (2008). Evaluating the process of a genetic algorithm to improve the back-propagation network: A Monte Carlo Study. *Expert Systems with Applications*. No.36(2). P.1459-1465.

Fariborz Y. Partovi & Murugan Anandarajan. (2008). Classifying inventory using an artificial neural network approach. *Computers & Industrial Engineering.* No.41(4). P.389-404.

Fumiaki Takeda & Sigeru Omatu. (1995). High Speed Paper Currency Recognition by Neural Networks. *IEEE Transactions on Neural Networks.* No.6(1). P.73-77.

Ge, Jike, Qiu, Yuhui, Wu, Chunming & Pu, Guolin. (2008). Summary of Genetic Algorithms Research. *Application Research of Computers.* No.10(10).

Jiansheng Wu & Mingzhe Liu. (2005). Improving generalization performance of artificial neural networks with genetic algorithms. *Proceedings of the IEEE International Conference on Granular Computing.* P.288-291.

Miller, J.A, Potter, W.D, Gandham, R.V & Lapena, C.N. (1993). An evaluation of local improvement operators for genetic algorithms. *IEEE Transactions on Systems, Man and Cybernetics.* No.23(5). P.1340-1351.

Ming Chen & Zhengwei Yao. (2008). Classification Techniques of Neural Networks Using Improved Genetic Algorithms. *Proceedings of the Second International Conference on Genetic and Evolutionary Computing.* P.115-119.

Min-Lun Lan, Shing-Tai Pan & Chih-Chin Lai. (2006). Using Genetic Algorithm to Improve the Performance of Speech Recognition Based on Artificial Neural Network. *Proceedings of the First International Conference on Innovative Computing, Information and Control.* P.527-530.

Omatu, S, Yoshioka, M & Kosaka, Y. (2007). Bank note classification using neural networks. *Proceedings of the IEEE Conference on Emerging Technologies & Factory Automation.* P.413-417.

Qiang Gao, Keyu Qi,Yaguo Lei & Zhengjia He. (2005). An Improved Genetic Algorithm and Its Application in Artificial Neural Network Training. *Proceedings of the Fifth International Conference on Information, Communications and Signal Processing.* P.357-360.

Qiuwen Zhang & Cheng Wang. (2008). Using Genetic Algorithm to Optimize Artificial Neural Network: A Case Study on Earthquake Prediction. *Proceedings of the Second International Conference on Genetic and Evolutionary Computing.* P.128-131.

Table 1. Measured data

Harmonic coefficient (a)	0.4	0.5	0.6	0.7
Identification rate (%)	98.38709	98.185458	99.395156	98.790323

Table 2. Measured data of identification rate

Identification method	BP	SGA+BP	IMGA+BP
Identification rate (%)	96.57	98.38	99.59

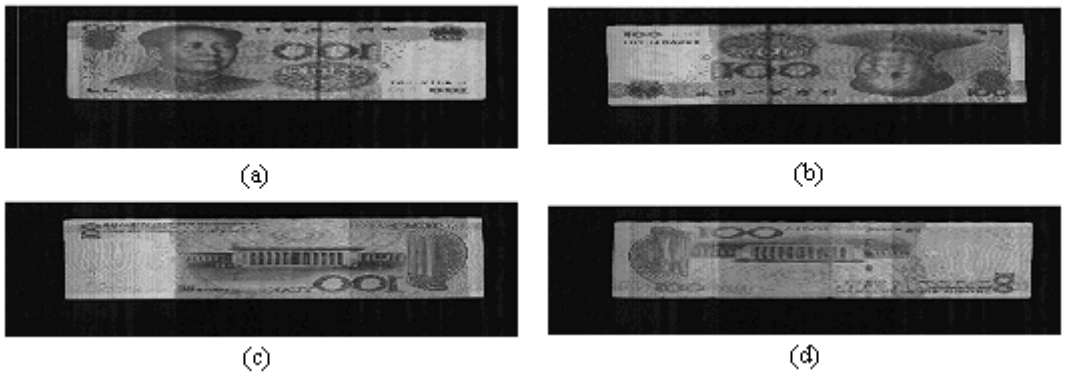

Figure 1. Four Faces of the Paper Money (a. the first face, b. the second face, c. the third face, d. the fourth face)

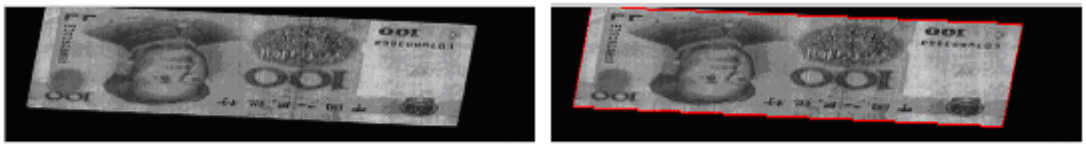

Figure 2. Paper Money Framing (a. original image, b. paper money framing)

Figure 3. Slope Correction Image (a. original image, b. slop correction image)

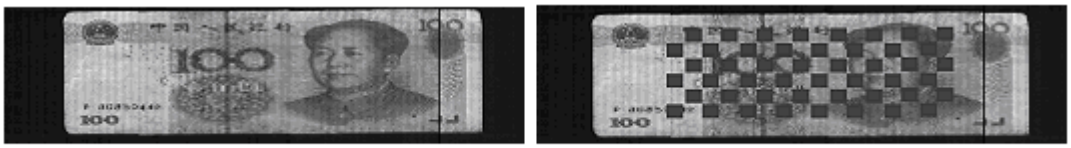

Figure 4. Paper Money Image Character Abstraction (a. original image, b. character image)

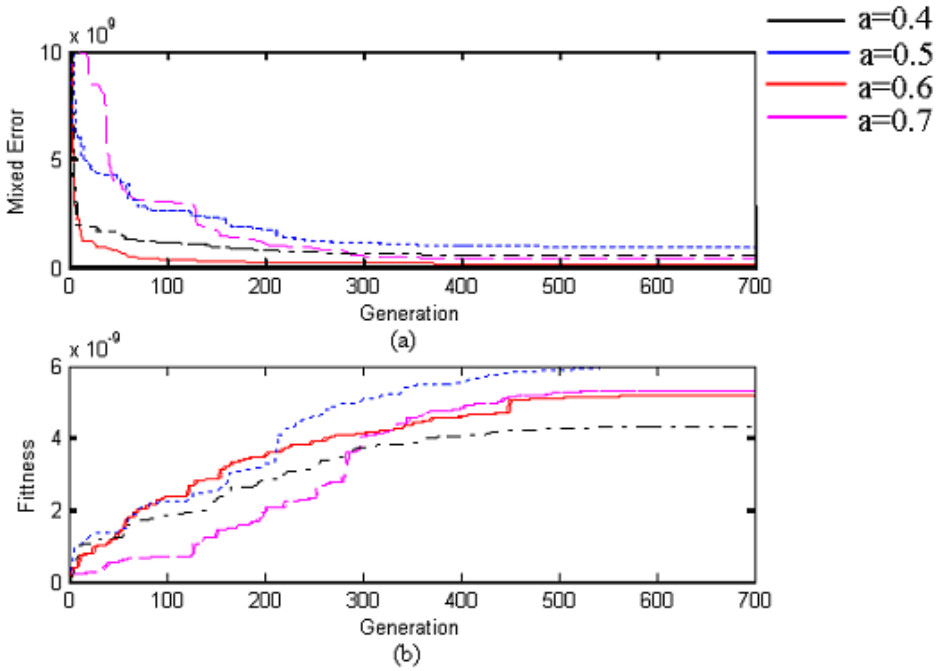

Figure 5. Fitness Function Error and Fitness Function Value Curves (a and b respectively are the fitness function error and the fitness function value curve when $\alpha = 0.4, 0.5, 0.6, 0.7$)

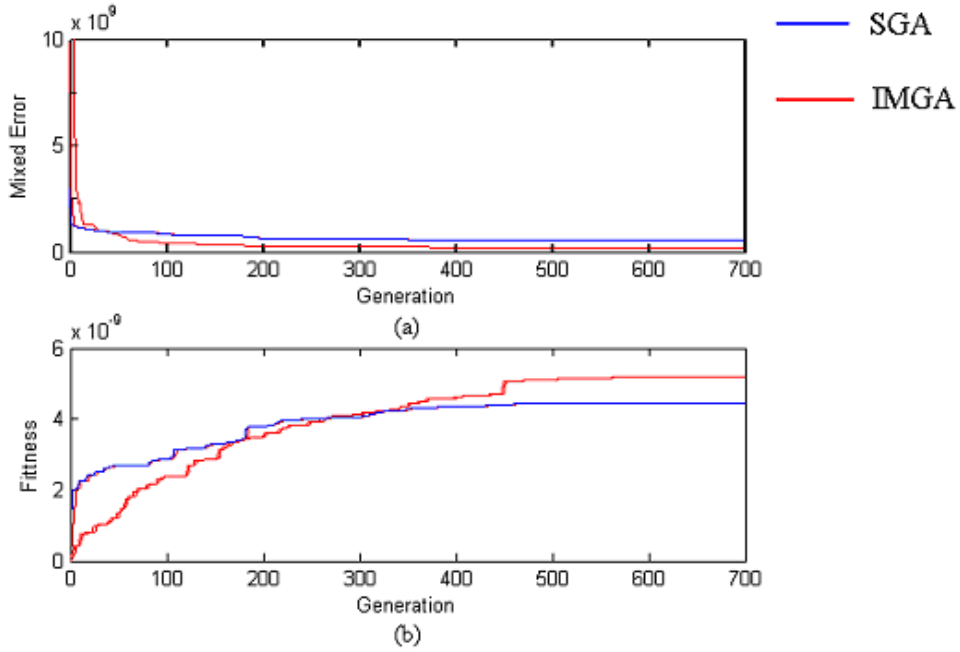

Figure 6. Fitness Function Error and Fitness Function Value Curves (a. SGA fitness function error and fitness value curve, b. IMGA fitness function error and fitness value curve)

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 2, No. 4

November 2009

Developing a Secure Web Application Using OWASP Guidelines

Khairul Anwar Sedek

Faculty of Computer Science and Mathematics, Universiti Teknologi MARA (UiTM)

Kampus Arau, 02600 Arau, Perlis, Malaysia

Tel: 60-19-474-6960 E-mail: khairulanwar@perlis.uitm.edu.my

Norlis Osman

Faculty of Computer Science and Mathematics, Universiti Teknologi MARA (UiTM)

Kampus Arau, 02600 Arau, Perlis, Malaysia

Tel: 60-19-477-2808 E-mail: norlis@perlis.uitm.edu.my

Mohd Nizam Osman

Faculty of Computer Science and Mathematics, Universiti Teknologi MARA (UiTM)

Kampus Arau, 02600 Arau, Perlis, Malaysia

Tel: 60-19-413-5362 E-mail: nizamos@perlis.uitm.edu.my

Hj. Kamaruzaman Jusoff (Corresponding author)

Faculty of Forestry, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

Tel: 60-3-8946-7176 E-mail: kjusoff@yahoo.com

The research is financed by Research Management Institute (RMI) of UiTM.

Abstract

Developing a secure Web application is very difficult task. Therefore developers need a guideline to help them to develop a secure Web application. Guideline can be used as a checklist for developer to achieve minimum standard of secure Web application. This study evaluates how good is OWASP guideline in helping developer to build secure Web application. The developed system is then tested using code auditing and penetration testing to identify the achievement of the system security for the application. After applying the testing techniques from Open Source Security Testing Methodology (OSSTMM) on the Top Ten Critical vulnerabilities as defined by OWASP, a standard measure score are calculated. The score is used to decide on the level of security of the developed web application. A high percentage score would indicate that the guideline helps in building a secured web application. Hence, the result proved that OWASP guideline is effective in ensuring the trustworthiness of the system and can be used as referral by other web developer especially in developing applications for a university.

Keywords: Web Application, Security

1. Introduction

In the age of Internet and World Wide Web, system security has become an important issue in any global web based information systems. This can be seen from the strong commitments of system security professionals, the research community, and major application software vendors. Recently web technology has developed rapidly and affected people in many aspects of lives and working. Many daily activities, which required face-to-face interaction, can now be conducted over the World Wide Web. Web applications are crucial components of our life. They cover critical activities such as economic transactions, e-commerce, e-government, e-business, e-procurement, e-education and many more.

The processes of building a secure web application need one or more guidelines to make it a secure system. Without guideline, it is impossible to develop a secure system. Gritzales & Spinelis (1997) provide the best practice for addressing security issues and threats, which can be prevented using security services. Stuart et al. (2001) has been addressing a very comprehensive guideline including system, network, and software security. Ed (2002) provided a step-by-step guide to computer attacks and effective defences including web application. Darothy (1998) assert that the best defence against security breaches is to make use of the tools and knowledge of good software engineering practice to prevent security attacks by developing and evolving secure system. This means that the requirements related to security issues must be identified and included early in the development and evolution of systems. Care must be taken to ensure that the security requirements are correct and complete.

The first OWASP (2003) issued the top 10 most critical web application security vulnerabilities to be considered in building secure web application with an update on the latest vulnerabilities in 2004. OWASP issued the latest Top 10 vulnerabilities (2007) which show that A1-Cross Site Scripting (XSS) has moved to the top of the list from 4th place and A2-Injection Flaws from 6th place to 2nd place. While A7-Broken Authentication and Session Management vulnerability has moved down from 3rd place in the list to the 7th placement. Several new vulnerabilities have been identified in the Top 10 2007 list such as A3-Malicious File Execution, A4-Insecure Direct Object Reference, A5-Cross Site Request Forgery (CSRF) and A9-Insecure Communications. Some vulnerability evolved from the old vulnerability into its own vulnerability definition, for example A10-Failure to Restrict URL Access is redefined from the previous A2 2004-Broken Access Control and A8-Insecure Cryptographic Storage is redefined from A8 2004 Insecure Storage. Meanwhile vulnerabilities such as unvalidated input, buffer overflow and denial of service have been taken out from the top 10 2004 list.

Mark et al. (2002) provided an open source document of guideline to building secure web application. These documents are intended to help developer to design, building and maintain a secure web application.

In the article "Buzzing About Security", Sandra (2002) explained why developer should have a framework as guideline to building secure web application. She recommends OWASP because it is an open source document where everybody can use it for developing, building, and testing secure system.

Pete (2002) produced an Open-Source Security Testing Methodology (OSTMM). It created an accepted method for performing a thorough security test including Internet presence points, information security, social engineering, networking, and physical security. Mark Curphey (2007) has produced a draft of OWASP Web Security Certification Criteria document to be used to test and certify the security of Web application. It can be a framework of Web application security certification. OSTMM has more comprehensive testing methodology to measure the result of security testing.

Andrew et al. (2003) have evaluated security features of Microsoft Windows Server 2003 with .Net Framework and IBM WebSphere. The study evaluated the level of effort required for developers and administrator to create and deploy secure web application. Both platforms provide infrastructure and effective tools for building secure application but the .Net platform scored higher than WebSphere.

Most of the studies discussed about security tools and how to build secure web application and also on current web threats. However, there is no evidence that the security tools or recommended security practice is adequate to build a secure web application. To conclude, there is no study on evaluation of security guideline or standard that can be followed by a developer in building a secure web application.

In this study, we evaluate how effective is OWASP guideline to help developer to develop secure Web application. To evaluate, OWASP guideline is used to develop secure Web application.

For this purpose, the required activities are integrated tightly into the development process. The security measures are carried out during the entire process, as early as possible when they become relevant. This ensures that security problem are discovered when they are still easy to counter. The process we used improves the quality (by its requirement on design, implementation and testing) and trustworthiness of the system and reduces evaluation time and cost.

2. Methodology

OWASP guideline is applied throughout the software development life cycle (SDLC) phases in application development which are system planning, system analysis, system design, implementation, and testing as shown in Figure 1. Security requirement defined in the OWASP such as authentication and authorization, input validation, and session handling is applied to ensure the system being developed is secure from security risks.

To have a standard measurement, score value for the vulnerabilities mentioned above is defined in Table 1 below. The table has been modified from the simplified web application framework to evaluate the guideline.

At the end of testing, all score will be summed up and the percentage will be calculated. This percentage will be analysed to determine whether the web application is secure or not. The following Table 2 represent the meaning of percentage in order to get the result or conclusion of the research for the guideline provided by OWASP.

3. Results and discussion

The study has successfully done 35 securities testing in the area of re-engineering, authentication, session management, input manipulation, output manipulation and information leakage testing. The test found 8 possible vulnerabilities out of thirty five possible testing (22.86%). The testing result is shown inTable 3.

The study has successfully done 35 securities testing in the area of re-engineering, authentication, session management, input manipulation, output manipulation and information leakage testing. The test found 8 possible vulnerabilities out of 35 possible testing (22.86%).

Based on the testing that we have done, for the area of re-engineering and information leakage security testing, with a result of 100%, we found that the guidelines helps immensely in building a secured web based application at least from the top 10 most critical vulnerabilities. Meanwhile testing the security in the area of authentication and session management, with a result of 78% and 76% respectively, shown the usage of guideline in this area gave adequate contribution to building a secured application. While in the area of input manipulation and output manipulation security, the above 85% result proved the guidelines to be considerable help in building a secured web application at least from the top 10 most critical vulnerabilities. Overall, the results of the security testing on ITMS yield the average security percentage of 86.27%.

Our study has demonstrated how we evaluated a Web application by using OWASP Guideline to building a secured Web Application. The guideline was evaluated using OSSTMM proposed by Pete Herzog, with the development for Industrial Training Management System (ITMS) Web application as a case study. This study has successfully applied the OWASP guideline to ITMS Web application. The result of all criteria that was evaluated indicated that OWASP contributed significantly in developing a secured Web application at least in reducing the number of security vulnerabilities especially for Web based university application.

4. Conclusion

The guideline was evaluated using OSSTMM proposed by Pete Herzog, with the development for Industrial Training Management System (ITMS) Web application as a case study. This study has successfully applied the OWASP guideline to ITMS Web application. The result of all criteria that was evaluated indicated that OWASP contributed significantly in developing a secured Web application at least in reducing the number of security vulnerabilities especially for Web based university application.

Overall, taking into account security does not make web design more complicated; it should be one of many natural elements of web design nowadays. It is not hard to consider if it is included into the process of web design right from the beginning.

Incomplete development processes leave the applications at risk, no matter how structured the company's development process may be. To achieve a greater level of application security, mature development practices that focus specifically on Web application security need to be implemented.

References

Andrew Jaquith, Frank Heidt, & Chris Wysopal, (2003). Security Evaluation: Microsoft Windows Server 2003 with .NET Framework and IBM WebSphere. Retrieved from http://www.atstake.com/research/reports/eval_ms_ibm/acrobat/atstake_eval_ms_ibm.pdf.

Andrew Jaquith. (2002). The Security of Appplications: Not All Are Created Equal. Retrieved from http://www.atstake.com/research/reports/acrobat/atstake_app_unequal.pdf.

Boiler. (2003). Hacking Techniques: Issue#2 – Boucing Atttacks. Retrieved from http://www.governmentsecurity.org/articles /HackingTechniquesIssue2-BouncingAttacks.php.

David Endler, Brute-Force Exploitation of Web Application Session IDs. Retrieved from http://www.cgisecurity.com /lib/SessionIDs.pdf

Dorothy E. Denning. (1998). Cyberspace Attacks and countermeasures". In Internet Besieged Countering Cyber Scofflaws, ACM Press, New York, N.Y.

Ed Skoudis. (2002). Counter Hack: A Step-by-Step Guide to Computer Attacks and Effective Defenses. Upper Sadle River, NJ: Prentice Hall PTR.

Grizalis, Stefanos & Spinellis, Diomidis. (1997). Addressing Threats and Security Issues in World Wide Web Technology. In Proceeding CMS '97 3rd IFIP TC6/TC11 ,33-46. IFIP, Chapman & Hall. Retrieved from http://www.dmst.aueb.gr/dds/pubs/conf/1997-CMS-WebSec/html/w3sec.html.

Mark Curphey. (2004). The OWASP Testing Project, (2004, December). Retrieved from http://www.owasp.org/index.php/OWASP_Testing_Project

Mark Curphey. (2004). The Ten Most Critical Web Application Security Vulnerabilities. Retrieved from http://www.owasp.org.

Mark Curphey. (2007). SpoC 007 - The OWASP Web Security Certification Framework. Retrieved from http://www.owasp.org/index.php/SpoC_007_-_The_OWASP_Web_Security_Certification_Framework.

OWASP Top Ten Web Application Vulnerabilities Version 1.0. (2003). Retrieved from http://prdownloads.sourceforge.net/OWASP/OWASPWebApplicationSecurityTopTen-Version1.pdf?download.

OWASP Top 10 2004. (2004). Retrieved from http://www.owasp.org/index.php/Top_10_2004.

OWASP Top 10 2007 (2007). Retrieved from http://www.owasp.org/index.php/Top_10_2007.

Pete Herzog. (2002), Open Source Security Testing Methodology Manual (OSSTMM). Retrieved from http://isecom.securenetltd.com/osstmm.en.2.1.pdf.

Sandra Kay Miller. (2002, January). Buzzing About Security. InfoSecurity. Retrieved from http://infosecuritymag.techtarget.com/2002/jan/departments_news.shtml.

Scott Berinato. (n.d.). The Bugs Stop Here. Retrieved from http://cio.idg.com.au/index.php?id=597539172.

Shynlie Simmons, Hacking Techniques: Web Application Security. (2005, November) East Caroline University. Retrieved from http://www.infosecwriters.com/text_resources/pdf/HackingTechniques_WebApplicationSecurity.pdf.

Stuart McClure, Joel Scambray, & George Kurtz, (2001). Hacking Exposed: Network Security Secrets and Solutions, Third Edition (3). : Osborn/McGraw Hill.

Table 1. Score value for the guideline evaluation

Score Value	Score Meaning
1	Not secure
2	Partly not secure
3	Fully secure.

Table 2. The definition of the security percentage calculated

Total Score	**Definition**
Less 25%	The guideline failed to help building a secure web application
26% - 50%	The guideline help eliminate some vulnerabilities but not enough to have secure application
51% - 79%	The usage of the guideline is adequate to build secure application
80% - 100%	The guideline helps building secure web application at least from the top 10 most critical vulnerabilities.

Table 3. Security Testing Result

No.	**Items**	**Score (1-3)**
	Re-Engineering	

1	Decompose or deconstruct the binary codes, if accessible	3
2	Determines the protocol specification of the server/client application	3
3	Guess program logic from the error/debug messages in the application output program behaviours/performance	3
	TOTAL	**9/9**
	Authentication	
4	Find possible brute force password guessing access points in the application	3
5	Find a valid login credentials with password grinding, if possible	1
6	By pass authentication system with spoofed tokens	3
7	By pass authentication system with replay authentication information.	1
8	Determine the application logic to maintain the authentication session – number of (consecutive) failure logins allowed, login timeout etc.	3
9	Determine the limitations of access control in the application – access permissions, login session duration, idle duration	3
	TOTAL	**14/18**
	Session Management	
10	Determine the session management information – number of concurrent session, IP-based, authentication, role-based authentication, identity based authentication, cookies usage, session ID in URL encoding string, session ID in hidden field variables, etc	1
11	Guess the session ID sequence and format	3
12	Determine the session ID is maintained with IP address information; check if the same session information can be retried & reused in another machine	1
13	Determine the session management limitations – bandwidth usages, file download/upload limitations, transaction limitation, etc	3
14	Gather excessive information with direct URL, direct instruction, action sequence jumping and/or pages skipping	3
15	Gather sensitive information with Man-in-the-Middle attacks	1

16	Inject excess/bogus information with Session-Hijacking techniques	3
17	Replay gathered information to fool the applications.	1
	TOTAL	**16/21**
	Input Manipulation	
18	Find the limitations of the defined variables and protocol payload – data length, data type, construct format, etc.	3
19	Use exceptionally long character-strings to find buffer overflow vulnerability in the applications	3
20	Concatenate commands in the input strings of the applications	2
21	Inject SQL language in the input strings of database-tired web applications	3
22	Examine "Cross-Site Scripting" in the web applications of the system	1
23	Examine unauthorized directory/file access with path/directory traversal in the input strings of the applications	3
24	Use specific URL-encoded string and/or Unicode-encode strings to bypass input validation mechanism of the applications	3
25	Execute remote commands through "Server Side Include"	3
26	Manipulate the session/persistent cookies to fool or modify the logic in the server-side web application.	2
27	Manipulate the (hidden) field variable in the HTML forms to fool or modify the logic in the server-side web application	3
28	Manipulate the "Referrer". "Host", etc. HTTP Protocol variables to fool or modify the logic in the server-side web applications.	3
29	Use illogical/illegal input to test the application error-handling routines and to find useful debug/error message from the applications.	3
	TOTAL	**32/36**
	Output Manipulation	
30	Retrieve valuable information stored in the cookies	2
31	Retrieve valuable information from the client application cache	3

32	Retrieve valuable information stored in the serialized objects.	3
33	Retrieve valuable information stored in the temporary files and objects	3
	TOTAL	**11/12**
	Information Leakage	
34	Find useful information in hidden field variables of the HTML forms and comments in the HTML documents	3
35	Examine the information contained in the application banners, usage instructions, welcome messages, farewell messages, application help messages, debug/error message, etc.	3
	TOTAL	**6/6**

Table 4. Summary of the Result Security Testing

Security Testing Category	Marks	Percentage (%)
Re-engineering	9/9	100%
Authentication	14/18	78%
Session Management	16/21	76%
Input Manipulation	32/36	89%
Output Manipulation	11/12	92%
Information Leakage	6/6	100%

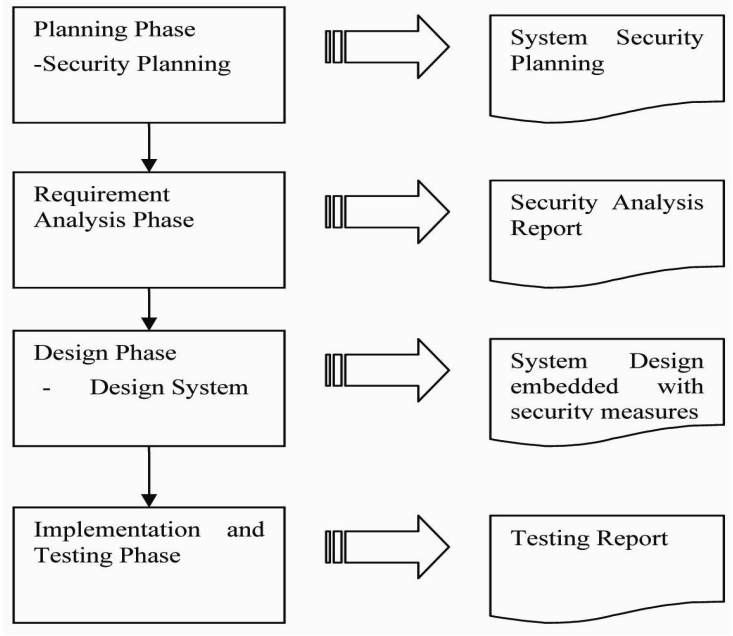

Figure 1. Security in Software Development Life Cycle

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 2, No. 4

November 2009

DOA Estimation for Coherent Sources in Transformed Space

Yuan Cui

Department of Computer, Chengdu Medical College

Chengdu 610083, China

E-mail: bubblecui@163.com

The research is financed by Asian Development Bank. No. 2006-A171(Sponsoring information)

Abstract

The existence of coherent sources results in the rank deficit of sample covariance matrix. Classic MUSIC(Multiple Signal Classification) can not classify coherent sources, and instead, generate an equivalent sources somewhere between them. In the proposed method, first, a specially designed transformation is constructed, which can suppress the coherent interfering sources while retain desired coherent sources. With the transformation the collected array signal can be mapping into a new data space. Since in the process of transformation, the contribution of the coherent interfering sources is suppressed, applying the classical music to the transformed data space will result in accurate DOA estimation of the coherent sources. Simulation experiment show that compared to the classical music, this method can accomplish accurate DOA estimation of coherent sources.

Keywords: China insurance industry, Foreign fund, Challenge

1. Introduction

Estimations of DOA by the means of sensor array processing is a hot spot and attracts many researchers to investigate it. It is widely applied in radar, sonar, seismology and underwater signal sources estimations. MUSIC is a class of high resolution DOA estimate algorithm and widely employed in those fields. MUSIC is known as a high resolution algorithm for DOA estimation, but in case of finite data samples it can not resolve adjacent sources with large power level differences between them. However, classical MUSIC has a serious drawback that it cannot conduct DOA with the existence of coherent sources, which results in rank deficit of source covariance matrix. It is well known that coherent sources extensively exist in real world, for example, multi-path propagation and jamming. To overcome this shortcoming, several effective methods have been developed, such as spatial smoothing and weighted subspace fitting.

Such array signal processing techniques also can be used to neuroscience. Here we also discuss how to use the method to coherent brain source localization.

One of the most active areas of research in contemporary neuroscience concerns the issue of functional connectivity and neuronal integration. At the microscopic level, increasing evidence show that relevant information in the brain is coded by accurate timing of neuronal discharges and Synchronized rhythmic neural firing has a role in solving the binding problem, i.e., the integration of distributed information into a unified representation. At all levels of description of cortical networks, the synchronization hypothesis get more and more supports in neurophysiologic literature. At the macroscopic level, Functional connectivity between cortical areas may appear as correlated time behavior of neural activity.

To investigate cortico-cortical synchrony noninvasively in the human brain, new analysis tools must be developed.FMRI have been used to estimates connectivity between brain areas. However, its temporal resolvability is not high enough to measure oscillatory activity and to observe transient formation of neuronal assemblies.

Magnetoencephalography (MEG) and electroencephalography (EEG) scalp recordings have unsuppressed temporal resolution to characterize neuronal coupling and commonly used to study inter-regional functional connectivity.

Indeed, task-dependent interactions have been reported between signals recorded by different MEG sensors or EEG electrodes. However, these findings are limited to correlations within the measurement device and reveal little on the synchrony between specific cortical areas.

The signal recorded by a MEG sensor or an EEG electrode cannot be directly attributed to the underlying cortical region. The complex relationship between the signal detected by a sensor and an activated brain area is given by the solution of

the forward problem (i.e., the calculation of the magnetic field or electric potential generated by a point source). Especially electric potentials (EEG) are smeared out because of the inhomogeneous conductivity structure of the human head.

The activity of even a small cortical area is recorded by several sensors, leading to severe spreading in sensor-based measures. The spreading is particularly problematic when describing interdependencies between signals.

Ideally, in order to study neuronal interactions one has to go beyond the sensor level, as need two steps: first sources have to been localized and then their temporal courses have to been estimated. Based on both the source locations and waveforms, one can investigate their interactions and psychophysiological implications. Many authors studied the algorithms for localizing neuronal sources. Among these methods, beamforming and music are two most popular algorithms and then attract many attentions.

Beamforming have been shown to provide reliable estimates both of the spatial location and the time courses of activity of neuronal sources. Furthermore, literature shows music is more accurate than beamforming. We select to develop music-type methods to localize the sources.

At mentioned above, Functional connectivity between cortical areas may appear as correlated time behavior of neural activity. To study interregional interactions within brain, methods focusing on handling correlated time courses should be developed.

However, in principal, classic music can not deal with correlated sources. Some authors developed some modified music to this case of weakly and moderately correlated sources. When sources is highly correlated (e.g. a extreme case, fully correlated), these methods will fail.

In 2001, J.Gross et.al. presented a pioneering technique, DICS, uses a spatial filter to localize coherent brain regions and provides the time courses of their activity. DICS is beamforming type method, designed for interactions between sources at specified frequency bands.

Nonetheless, there is a pitfall (16)in this approach that meg beamformer methodology is based on the underlying assumption that no distinct neuronal sources are perfectly linearly related. In fact, in the presence of high, long-lasting, source correlation, the estimated signal intensity and temporal distortion will grow worse/deteriorate. Furthermore, DICS need to select reference points before imaging of the coherent sources, as may results in different operators get different results.

Here, we present a subspace-based method to localize fully correlated sources (the correlation coefficient between sources equal to 1). Throughout the paper the terms coherent sources will be used to denote such strictly linear relationships between time courses. The key point of this method is to decrease correlation between sources largely enough that classic music can easily localize them. After accurately finding the positions of coherent sources, estimations of source time courses is relatively easy and have many reliable methods to do that. Since the present method by adding an inverse source into the head model classify coherent sources.

This paper we focus on a new DOA estimate method, which can not only estimate DOA of coherent sources, but also can identify closely paced sources. In the following analysis, for the simplicity, we only discuss the case of two sources. The case of more than two sources is more complex, but still can be deled with by this means.

2. Methods

2.1 recall of classical MUSIC

Classical MUSIC, initially proposed by Schmidt, is used to solve the problem of DOA estimation in array signal processing. Suppose there are r sources impinging on m array sensors from different scalar directions. The manifold vector may therefore be specified as $\boldsymbol{a}(\boldsymbol{\theta})$. The set of r manifold vectors may been expressed as

$$\boldsymbol{A}(q) = [\boldsymbol{a}(q_1),...,\boldsymbol{a}(q_r)] \tag{0.29}$$

The data sample x(t) collected from array sensors can be expressed as

$$\boldsymbol{x}(\boldsymbol{t}) = \boldsymbol{A}(q)\boldsymbol{s}(\boldsymbol{t}) + \boldsymbol{n}(\boldsymbol{t}) \tag{0.30}$$

Where s(t) is the time courses of sources and n(t) is Gaussian noise. Based on the assumption that the additive noise n(t) are uncorrelated with the source time courses, then,

$$\boldsymbol{E}\{\boldsymbol{n}(\boldsymbol{t})\boldsymbol{n}^{\boldsymbol{H}}(\boldsymbol{t})\} = s^2\boldsymbol{I} \tag{0.31}$$

Where superscript H denotes the Hermitian transpose. The autocorrelation of x(t) can be partitioned as

$$\begin{aligned} \boldsymbol{R} &= \boldsymbol{E}\{\boldsymbol{x}(\boldsymbol{t})\boldsymbol{x}^{\boldsymbol{H}}(\boldsymbol{t})\} \\ &= \boldsymbol{A}(\mathrm{Q})(\boldsymbol{E}\{\boldsymbol{s}(\boldsymbol{t})\boldsymbol{s}^{\boldsymbol{H}}(\boldsymbol{t})\})\boldsymbol{A}^{\boldsymbol{H}}(\mathrm{Q}) + s^2\boldsymbol{I} \\ &= \mathrm{F}[\mathrm{L} + s^2\boldsymbol{I}]\mathrm{F}^{\boldsymbol{H}} = \mathrm{F}_s\mathrm{L}_s\mathrm{F}_s^{\boldsymbol{H}} + \mathrm{F}_n\mathrm{L}_n\mathrm{F}_n^{\boldsymbol{H}} \end{aligned} \tag{0.32}$$

Where Φ_s, Φ_n are the signal subspace and the noise subspace, respectively. since the signal subspace is orthogonal to the noise subspace, we can obtain the cost function as the following to estimate the DOA,

$$J_{mu} = \frac{1}{a_i^h(q)\mathrm{F}_N\mathrm{F}_N^H a_i(q)} \qquad i = 1,...,N \tag{0.33}$$

Theoretically, while $a_i(\theta)$ exactly is the manifold vector associated with the actual sources, $a_i^h(q)\mathrm{F}_N\mathrm{F}_N^H a_i(q) = 0$. The MUSIC algorithm uses this property to estimate direction of sources. when applied in real data, because of the effects of noise and computation errors, J does not equal to zero. $a_i(\theta)$, which let J reach to its local maximums of J, is the manifold vector of the true sources. by this means, one can find the directions of the true sources.

2.2 MUSIC's disability of coherent source estimation

Subspace based approaches have two advantages 1) decrease computational load 2) avoid nonlinear search. However, it is based on an assumption that the source time courses is independent or weak correlated. To strongly correlated sources, classic music doesn't not handle them. Modified music has been developed to deal with this case. R music Rap music, fines et al can deal with strongly correlated sources. However, when sources are fully correlated (the Correlation coefficient between sources proximally equals 1), all those methods will fail. For instance, the time course of source 1 is s_1, and source 2 s_2. Suppose $S_2=kS_1$. Their lead matrix is a_1 and a_2, respectively. Then the scalp EEG $y=a_1s_1+a_2s_2$. Since $s_2=ks_1$, $y=(a_1+ka_2)s_1$. In theory, instead of the true locations of both s_1 and s_2 which are associated with lead matrix s_1 and s_2, classic MUSIC will mistakenly localize s1 and s2 at the location corresponding to the lead matrix a1+ka2. Our idea is, since the cc is 1, if cc can decrease to a small enough degree that classic music can identify them, we can easily identify them with any further processing. But the new problem arises, it is an inverse problem and we have no prior information of the source position. Sitting the constructing source at each gird of the whole head model is a kind of methods. While constructing source is just positioned at the location of any true sources, since the cc between the combination of constructing source and one of true source and the other is small enough, the cost function will find two peaks and classic music will easily localize them. Sitting construct source at other location results in the cost function getting one peak since constructing source and two true sources are coherent and then they will only generate an equivalent source. From the number of local peaks, one can know when the constructing source is just the positions of the true sources.

2.3 DOA estimation for coherent sources in transformed space

2.3.1 the algorithm formulation

Firstly, with a prior information, the approximate direction of the coherent interfering sources can be estimated. The collection of these direction vectors, $a(r_i) \quad i = 1,\cdots,N$, can be represented as

$$H = [a(r_1) \quad \cdots \quad a(r_N)] \tag{0.34}$$

We can construct a transformation matrix G,

$$H = [a(r_1) \quad \cdots \quad a(r_N)] \tag{0.35}$$

$$G = I - H(H^T H)^{-1} H^T \tag{0.36}$$

Applying this transformation to the collected array signal, we can get,

$$y(t) = Hx(t) \tag{0.37}$$

Then, In transformed space, the signal from the direction to be suppressed can be obtained,

$$x(t) = G(HS) \tag{0.38}$$

Since $GH = [0]$ by injecting equation (1,9) into (1.11), the signal coming form coherent interfering direction will be suppressed. And thus, its effect on coherent sources is removed and can be estimated correctly.

2.3.2 Simulation test

In order to validate the effectiveness of the proposed method, we take a conventional two source uniform linear array as examples to compare this method with the other sequential forms. We follow the simulations in in order to draw performance comparison between the various sequential forms of MUSIC. The sources are far field narrowband and impinging on the array from scalar direction θ. The array manifold vector may therefore be specified as

$$a(\theta) = [1, e^{j\pi\sin\theta}, ..., e^{j\pi(m-1)\sin\theta}]^T \tag{0.39}$$

Where θ=0 is broadside to the array, and $\|a(q)\| = m$. The source time series are assumed to be complex zero mean Gaussian distribution with covariance matrix P. suppose there are 15 sensors and two sources at 25 and 30 degree. In another simulation, the angles of two sources are at 14 and 16. The source covariance matrix is specified as

$$P = \begin{bmatrix} 1 & \gamma \\ \gamma & 1 \end{bmatrix} \tag{0.40}$$

Where γ determine the degree of correlation between these two sources with equal power. the variance of noise is set to unity, such that the ratio of signal to noise is also unity.

Simulation experiment 1: We set γ to 1, that is, the two sources are completely coherent. Fig. 1. shows the results obtained by the conventional MUSIC. Obviously, the conventional MUSIC can not identify the directions of the two coherent sources correctly, but place an equivalent false sources somewhere, about 28 degree, between them.

Simulation experiment 2: In order to suppress the energy from the directions ranging from 23 to 28 degree, according to equation (1.8), we construct a transformation matrix G. Using G to transform the collected array signal, we can get the correct estimation of source 2. The result is showed in Fig.1, in which, we can see the spectrum reaches peak at direction 30 degree.

Since the DOA of source 2 was estimated correctly, we can design a transformation matrix to suppress the energy from source 2 and achieve DOA estimation of source 1(Fig.1).

From the simulation, we can conclude that for coherent sources, the method have better performance and estimate directions of sources with less error. While the conventional MUSIC encounters difficulty in the presence of completely coherent sources.

3. Conclusions

The paper presented a new DOA estimate method for coherent sources. Preliminary simulation test confirm the effectiveness of this method. In contrast to the other methods, the proposed method has some advantage. To further study and improve this method is our next step works.

References

DeGroat, RD, Dowling, EM and Linebarger, DA. (1993). The constrained MUSIC problem. Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, *IEEE Transactions on*], 1993. 41(3): p.1445-1449.

Liu Gang, Lu Xinhua and Xiao Yang, (2006). MUSIC based spatial spectrum estimation algorithm in array signal processing, microcomputer information, 2006(04): p. 302-303, 292.

Mosher, J.C. and R.M. Leahy. (1998). Recursive MUSIC: a framework for EEG and MEG source localization. *IEEE Trans Biomed Eng,* 1998. 45(11): p. 1342-54.

Mosher, J.C. and R.M. Leahy. (1999). Source Localization Using Recursively Applied and Projected (RAP) MUSIC. *IEEE Trans Biomed Eng*, 1999. 47(2): p. 332-340.

Schmidt, R.O. (1986). Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 1986. AP-34(3): p. 276-280.

Stoica, P., P. Handel, and A. Nehoral. (1995). Improved sequential MUSIC. Aerospace and Electronic Systems, *IEEE Transactions on*, 1995. 31(4): p. 1230-1239.

Computer and Information Science
www.ccsenet.org/journal.html
Vol. 2, No. 4
November 2009

Construction of Information Disaster Recovery for Hospitals

Juan Xu
School of Information, Yunnan University of Finance and Economics, Yunnan 650221, China
E-mail: Xujuan@ynufe.edu.cn

Abstract

When data disaster happens, the disaster recovery system (DRS) can respond automatically and real-time, and quickly restart the application system in the redundant device to keep the processing of the business. The construction principle of the disaster recovery system has been studied, and combining the base function and national standard of the hospital information system (HIS), the data disaster recovery mode in the hospital information system was analyzed systematically and the constructive project was proposed in the article.

Keywords: Hospital information system, Data disaster recovery, Business recovery, Remote backup

1. Introduction

With the quick development of the computer technology, the information technology has been extensively applied in the medial and heath industry. On the one hand, China is actively pushing the requirement and standard of the hospital informationization construction. On the other hand, in the informationization construction of hospital, the management quality of hospital has been enhanced, and hospitals have benefited much. And hospitals have also been trying to push the informationization construction all along.

The medical information system has been applied in hospitals widely, which makes hospitals acquire notable management benefit and economic benefit. When hospitals enjoy quick service decision-making and convenient management because of informationization, they are in the danger of data loss (Dong, 2001, P.66-68).

Danger one: The deep development of hospital informationization, the patients continually added, and the business mode of 24×7 every weak make the data of HIS increase by the class of TB, and the explosive data need to continually add storage devices, and the storage and backup system in HIS is always approaching to breakdown.

Danger two: When the computer system encounters natural disasters, computer crime, computer virus, hardware/software error and man-made mistake operations, how to guarantee the security of users' data and the continual operation of the information management system.

Danger three: The upgrade of the old system will certainly bring the problem of data transfer, and how to ensure that the system doesn't lose key data in the transfer process?

Danger four: The invalidation of the key nodes in the network system will induce the stop of the data operation. How to quickly respond the faults and start the data recovery system to continue the business processing?

The virus attack and the loss of backup data will induce death warrant to the continual operation of hospital. The urgent problems how to guarantee the normal running of the system and prevent the data loss because of faults or disasters should be solved by HIS as soon as possible. If the security of the hospital data can not be guaranteed, the meanings of large numbers of network investment will be lost.

2. Hospital information disaster recovery system

2.1 Hospital information system

HIS (Informationization Work Lead Group Office of Chinese Ministry of Public Health, 2001) means to utilize modernization measures such as computer software and hardware technologies and network communication technology to comprehensively manage the patient flow, logistics and financial flow of the hospital and various departments, and collect, store, process, abstract, transport, gather and generation various information generated in various stages of the medial activity, and provide comprehensive and automatic management and the information system of various services for the whole running of hospital. HIS is the necessary infrastructure and support environment in the construction of modern hospital.

HIS belongs to the most complex type in the enterprise class information system, and it is decided by the objective, task and character of hospital. It can not only follow and manage the management information generating in the patient flow, financial flow and logistics with other management information systems (MIS) to enhance the running efficiency of the

whole system, but also support the whole medical, teaching and researching activity taking the patient medial information record as the center.

2.2 Disaster recovery system

Disaster recovery (DR) (Zhang, 2004) is a concept with extensive category, and generally, all contents relative to the operation continuity should be brought into the disaster recovery. DR is a systematic engineering, and it includes all aspects supporting the user operation. For IT, DR is the computer system to prevent that the user operation system suffers various disasters. And DR is also represented as a kind of initiative taking precautions, and it is not "taking precautions after suffering a loss" after disaster happens.

From the strict view, the DR usually talked by us means that when the production stations are destroyed by the disaster, other redundant stations established by the user can replace the normal operation and keep the continual operation. To achieve higher usability, many users even establish multiple redundant stations.

To prevent above possible disasters and reduce possible losses furthest, the disaster recovery system (DRS) is often established for pivotal operations (Xie, 2004). The establishment of DRS needs two parts, i.e. the data disaster recovery and the application disaster recovery. The data disaster recovery means to establish a distant data system which is a real-time copy of the local key application data. The application disaster recovery is to establish a set of complete copy application system corresponding with the local production system in the different place, and in the disaster, the remote system could quickly replace the local system. The data disaster recovery is the guarantee to fight disaster, and the application disaster recovery is the construction target of the disaster recovery system.

Technically, there are two main indexes, RPO (Recovery Point Object) and RTO (Recovery Time Object) to measure the disaster recovery system (IBM, 2005), and the RPO presents the data quantity allowed to be lost when the disaster happens, and RTO presents the recovery time of the system. When RPO and RTO are smaller, the usability of the system is higher and user's investments will be larger. Of course, the class of the disaster recovery system is also decided by the protection class and the significance of the operation application, and the construction should be based on the effective capital utilization and the existing system rebuilding. RPO and RTO must be confirmed by different operation demands after the risk analysis and operation influence analysis are performed. To different operations, the demands of RPO and RTO are also different.

3. Disaster recovery mode of HIS

Disasters can not be predicted basically, and the loss also can not be estimated exactly, but the influence of the disaster always is deathful for the subsequent works of the hospital which has begun the informationization construction. The hospital data disaster recovery is one part of the continual plan of hospital operation, and it is most important to establish HDDRS. When the data disaster happens, the disaster recovery system (DRS) can respond automatically and real-time, and quickly restart the application system in the redundant device to keep the processing of the business.

According to the actual situation of hospitals in China, the information system of hospital mainly includes HIS, LIS and PACA. And most hospitals only have HIS and LIS at present, and the data quantity they generates every year is not large, less than 10G. But the data generated by PACS are egregious, and it will generate more than 1TB data, so it is more important to protect the security of data.

For the hospital data recovery, the storage is the base, the backup is the core and the recovery is the key.

The data storage can not ensure the continual running of key business, and the data backup is necessary in the network of the enterprise class. The data backup can quickly recovery the back-up data, largely reduce the time of service interrupt, and provide good data service for users when the system is damaged or the data is lost.

Because the particularity of the hospital operation system, to ensure the continual running of the system, the server control center generally adopts the double-computer fault tolerance and the server cluster, or the cold backup computer to deal with the downtime to reduce the time of halt. The data backup in most hospitals is not perfect. For the hospitals without PACA, most of them only adopt the self-backup program of the database software to copy the data to the hard disk one time or two times one day, and only few of them adopt the magnetic tape backup and hard disk backup. For the large-sized hospitals which have used the PACS system, most of them adopt the LAN-free backup mode based on SAN to connect the LTO tape base in the fiber exchanger, and realize the automatic classified storage and backup by the backup software, which can completely release the bandwidth of the network. Both the magnetic tape (base) backup and the hard disk backup respectively have their corresponding advantages and disadvantages (He, 2004, P.41-45).

Only the local data backup can not fulfill the requirement of the hospital informationization construction. So the remote disaster recovery system must be established to protect the key data in the HIS. The key data will be copied to the remote disaster recovery center by means of the remote data backup technology. The remote disaster recovery center can monitor the activities in the production center and the local backup center real time, and once the faults occur in the local production center and the backup center, the operation process will be quickly replaced to the remote disaster

recovery center.

Combining present analysis, according to the operation demand and operation data scale of hospital, three-class modes are adopted to construct the hospital information data disaster recovery solution.

The fist class: Establishing stable production data storage center.

The second class: Establishing safe local data backup system.

The third class: Establishing remote disaster recovery system with quick response.

3.1 System structure of the HIS disaster recovery mode

The system structure of the HIS disaster recovery mode is seen in Figure 1.

3.2 To establish stable production data storage center

According to the operation contents and data quantity processed in HIS, the storage solution of the production data storage center can be established. The information service is composed by the application system N+1 cluster system, and the storage part is composed by the storage solutions based on iSCSI. The solution project of the production data storage center is seen in Figure 2.

PACA is the comprehensive application system to collect, store, manage, diagnose and process the digital medical hospital image information generated by the digital medical equipments such as CT, MR, US, X-ray apparatus, DSA and CR in the hospital, and it is one of most important parts of HIS. For large numbers of image information and accumulated data generated every day, the N+1 cluster can provide all application information service of HIS. According to the types of the collected information, five applied information servers and one backup computer are designed. Five application information servers provide the services respectively for the PACS, outpatient service, being-in-hospital, medicine, and material equipment and financial information.

iSCSI (small-sized computer system interface of Internet SCSI) is a standard to transfer data mass on the Internet or Ethernet. It was initiated by Cisco and IBM, and largely supported by the people who advocated the IP storage technology. And it also is a SCSI instruction set which can be running on the IP and be used for hardware device. Simply speaking, iSCSI can realize the running of SCSI agreement in the IP network, which can make it select the route on the gigabit Ethernet with high speed. The main function of iSCSI is to perform the encapsulation and reliable transfer of large numbers of data between the master computer system (initiator) and the storage device (target) on the TCP/IP network. In addition, iSSCI also can provide the encapsulating SCSI order in IP network and run on the TCP. iSSCI is the technical standard based on IP, and it can realize the connection between SCSI and TCP/IP, and for the users taking the LAN as the network environment, a few investments can help them to realize convenient and quick interactive information and data transfer and management.

3.3 To establish safe local data backup system

The local data backup can adopt the three-class backup. The first class backup is the hot backup of data, i.e. adopting the copy software to realize the synchronization of source-data and objective data. Each data updating operation is implemented on the production center and the local backup center at the same time. The second class backup is the code backup of data. Any technology has its own limitation. Though the copy software can realize high-level data protection, and it can protect data and realize the re-synchronization in time when the chain is in fault, or the main-array/assistant array is in the unattainable state or suffers natural or mechanical damage, but if the legal operation of source-data will induce the invalidation of database, the database of the objective data will be invalidated in the same way. So the cold backup of data can be adopted for the data source, for example, performing time added backup in the night of every Saturday. This project can provide the data protection to the man-made and application mistakes. The third class backup is the warm backup of data, i.e. the database copy technology. The complete data copy is reserved in the local backup center, and the updating log is transferred periodically to the local backup center by the production center through the network.

Generally, the disaster recovery system needs much investment, but the use probability is low, so the total cost of ownership (TCO) and the return on investment (ROT) should be seriously analyzed and computed, and in the local data backup system, one backup server is adopted as the backup system connecting with the tape base.

3.4 To establish remote disaster recovery system with quick response

In the various IT systems of enterprise, the production center is very important, and it always matches with a remote backup center. In the interior of the production center, various data protections have been implemented. When the fire or earthquake happens and the production center is in paralysis, the backup center will replace the production and continue to provide the network service. In the remote disaster recovery solution project, the remote disaster recovery center is established based on the iSCSI cluster system and it can actualize the uniform backup and system disaster recovery of the operation system.

Simply speaking, iSCSI is to encapsulate SCSI by TCP/IP and transfer it in the Ethernet. The high-speed gigabit iSCSI combines SCSI, Ethernet and TCP/IP.

The technology of iSCSI is mainly used to solve the remote storage problem (He, 2004, P.41-45).

3.4.1 To realize the data exchange among different places

Both different places have their own storage networks based on fiber (SAN), and the cost that two networks are connected by the fiber to realize the data exchange between two different places is too expensive. iSCSI is based on IP, and it can contain all parts in IP network, and if FC is converted into the data of IP, these data can be transferred by the traditional IP network, which will solve the problem of remote transfer, and when the data arrive at the other end, the data of IP are converted to the local FC storage network, so two fiber networks can be connect under low cost investment by iSCSI to realize the data exchange among different places.

3.4.2 To realize the data backup and disaster recovery among different places

By iSCSI, users can span standard Ethernet cable to establish actual SAN network at any place, and they need not to require special fiber channel network to transfer data between the server and the storage devices. iSCSI makes the remote mirroring and backup become possible, because without the distance limitation of fiber channel, the standard TCP/IP can make the data to transfer in the Ethernet. But from the view of data transfer, most iSCSI network transfer bandwidths are 1Gbit at present, and if the FDX is realized, the bandwidth can achieve 2Gbit, and the bandwidth of the second generation product can also achieve 2Gbit, and in the future third generation general iSCSI standard, the bandwidth will achieve 10Gb, and to establish the remote disaster system by iSCSI will be easily realized.

4. Conclusions and expectations

By the disaster recovery system, the medical information system can achieve high usability, high security, high efficiency, high expansibility and high management property. As viewed from the operation and application layers, the disaster recovery processing and high usability of the data center can enhance the efficiency of service and increase users' satisfactions and competitive forces through ensuring continual 24-hours key operations.

The disaster backup has gradually turned from original tape-backup technology to the disk mirroring technology and from single-computer backup to the network backup, and the backup data center has gradually turned to the hot backup from cold backup, and the requirements of disaster recovery class are high and higher. At present, to construct continually useful system and ensure the sustainable operation of the operation system, and better provide services to users is the objective pursued by the hospital informationization construction, and it is the development direction of future disaster recovery to establish the data center without data loss, which can automatically perform the switch when the disaster happens to ensure the continual usability of the operation system through the disaster recovery, especially the remote application class disaster recovery. But the research about the remote application class disaster recovery is still in the start stage, and relative technology and documents are rare, and the implementation is very difficult. But its importance can not be ignored, and it is the base to construct the usability system with continual operation, and the development direction of future disaster recovery, and a set of complete technical theory is needed to support it at present.

References

Chen, Qiang, Ma, Liya & Zhao, Wei. (2006). Digitized Hospital Standard System Construction and Question. *Medical Information*. No.19(1). P.27-29.

Dong, Weiyuan & Wang, Mingbao. (2001). Enterprise Disaster Recovery from "911". *PC World China.* No.6(1). P.66-68.

He, Suining. (2004). Study on the Disaster Recovery Technology. *Modern Electronic Engineering*. No.4. P.41-45.

IBM. (2005). White Book of Disaster Recovery. [Online] Available: http://www.ibm.com.

Informationization Work Lead Group Office of Chinese Ministry of Public Health. (2001). *Basic Function Standard of Hospital Information System.* March of 2001.

Xie, Changsheng, Han, Desheng, Li, Huaiyang & Cao, Qiang. (2004). Grade and Technology of Data Disaster Recovery and Copy. [Online] Available: www.csip.cn/new/st/al/2004/0730/342.htm. (July 30, 2007).

Zhang, Feng. (2004). Expert Talking of Storage Technology: Disaster Recovery Builds a Port. [Online] Available: www.dostor.com/info/netstor/2004-10-29/0001921045. (Oct 29, 2004).

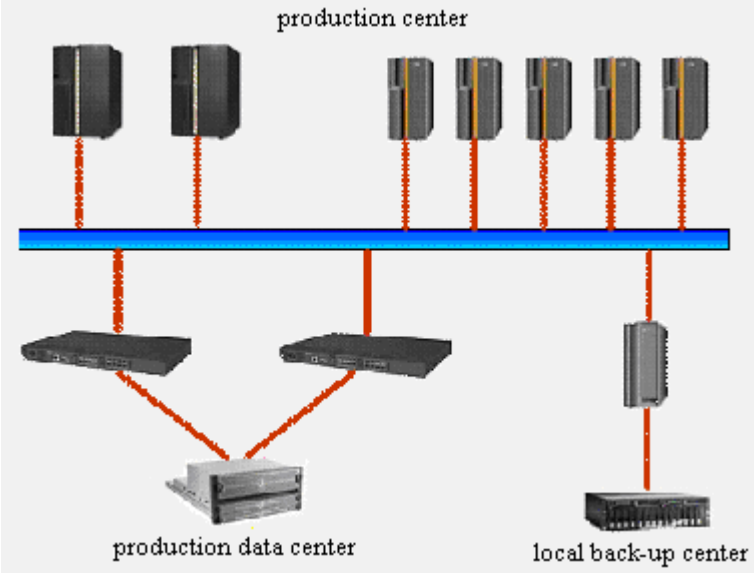

Figure 1. System Structure of the Hospital Data Disaster Recovery Mode

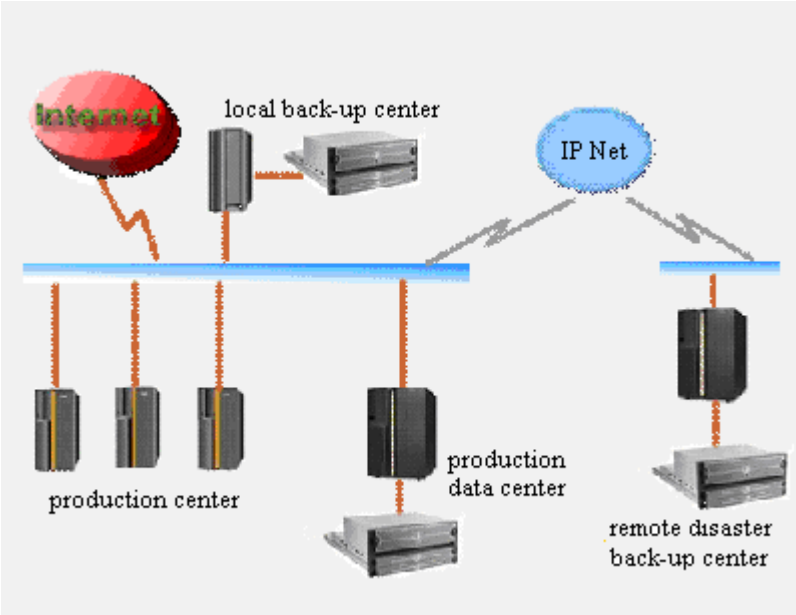

Figure 2. Solution Project of Production Data Storage Center

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 2, No. 4

November 2009

The Pastry Algorithm Based on DHT

Jihong Song (Corresponding author)

School of Information, Shenyang University of Technology, Shenyang 110178, China

Shaopeng Wang

School of Information, Shenyang University of Technology, Shenyang 110178, China

E-mail: wangshaopeng1984@163.com

Abstract

In the P2P network, how to quickly and accurately positioning of resources is a key measure of the performance. Nowadays, distributed P2P system generally adopts DHT search method, DHT-based P2P network search algorithm of P2P is a hot topic. The paper introduces the background and current research status of the DHT. Then focused on the Pastry algorithm, Finally Pastry algorithm was analyzed and some disadvantage about it was discussed.

Keywords: Pastry algorithm, DHT (Distributed Hash Table), Structured P2P

1. Introduction

P2P mechanism depends on the design and accomplishment of distributed system, where each system has nearly the same function and mission. These systems must assort with each other and not related to the central control. P2P system has the character of centralization. As far as the conventional storage and search strategy is concerned, all the position information needed by the central like server is stored in a single server. Therefore, its safety and robustness is weak. Flood like search has increased the communication burden, and therefore its inquiry function is not extensible when the system scale is increased. Moreover, as the inquiries are limited to a specific context, and therefore they can not be guaranteed that the network will be able to find the existence of the purpose of data.

In recent years, a great deal of research work has been done by many research team in the design of scalable search mechanism, e.g., they put forward a Chord, Pastry, CAN, etc. has been used to build a structured P2P system, distributed hash table (Distributed Hash Table, DHT). In this paper, we will introduce the basic principles of DHT, and also analyze the Pastry routing algorithm system.

2. The basic principle of DHT

In the structured P2P network, the client terminal is known as node, and data project is known as the object data. Namespace refers to the name domain system. The entire name in the domain name is unique. Name is used to mark the node. A typical name is its IP. Identifier refers to a unique integer in namespace, in the P2P system, it may be getting by the name of a node subsequently, keyword is a unique object identifier, object name available through a hash. In the DHT-based P2P systems, files are linked to keywords. Each document index is expressed as a (K, V) pair, K is called keywords, can be a file name (or description of document) of the hash value, V is the actual file storage node IP address (or description of the other nodes). The entire Index (that is, all (K, V) of) constitutes a large file index hash table. We just need to input the value of K, and then the destination file can be found from this table all the storage node address of the document. Then dividing the hash table into many small pieces to these small local hash table of the system in all the participating nodes, each node makes them responsible for the maintenance of a piece region. When do document inquiry, as long as the routing inquiries send the appropriate message to the node (the node to maintain the hash table contains a block to find the (K, V) on). There remains a problem, that is, nodes should be in accordance with certain rules to partition the overall hash table, which also determines the node's neighbors to maintain a specific node, so that routing can be carried out smoothly. Specific systems are also different rules, CAN, Chord, Pastry has its own rules, and it shows different characteristics.

3. Pastry Algorithm

Distributed routing system Pastry is proposed by Rostron and Druschel in 2001. Pastry is similar to Chord, the main goal is to create a completely centered, structured P2P systems, which can be efficiently target positioning, message routing to be efficient. Pastry identifier space is not organized into such as the Chord ring, but routing based on the approach of numerical identifiers.

3.1 The design of Pastry Algorithm

In Pastry, the nodes and data items are uniquely connected to the L bit identifier, i.e., the integer in the range of 0 ~ $^{L^2}$ -1 between the integer (L is typically 128). In such kinds of correlation, the identifier is corresponding known as a node ID or a keyword. The Pastry identifier is as the number of $^{2^b}$ -1 based series, in which b is typically 4. The value of the Keyword is located in the most recent node ID. In the above Figure, a Pastry identifier space is plotted, it has a 4bit identifier and b = 2. And therefore all the number has a base 4. The closest node to the keywords, such as keyword K01 is N01, node N10 which is located on the K03. K22 keywords have the same distance to nodes N21 and N2, so these two are held by the keyword node.

3.2 Router of Pastry

3.2.1 Router information.

Pastry node state can be divided into three main components. Chord routing table is similar to the target table space for storing the connection identifier; leaf collection is included in the marking of space in a similar line of nodes (like follow-up table in Chord); on the network in terms of localized nodes together in neighbors are listed in the collection.

(1). the routing y is assumed to be formed by N nodes, it has $\log_{2^b} N$ rows, and each row has $^{2^b}$ -1 entrance (b is a collocation parameter; the typical value is 1, 2, 3, 4). The entrance of N-th row points to the node NodeID. They share the $^{2^b}$ -1 table, with the first n bit the same, but the (n+1)th is different. The selection of value b is related to the length of router and the equivalence of router jumper. The larger is value b, the smaller is the jumper of router.

Meanwhile, more routing information is needed to maintain. The shadow of each line item in the routing table is corresponding to the current bit node number.

(2). Leaf collection: leaf node is stored in the collection nearest in value from the view point of the current node identifier. Collection of leaf nodes is needed in the routing information. It is obtained by taking the lower integer of $^{|L|/2}$, (the allocation value of $^{|L|}$ is $^{2^b}$). Taking the lower integer for NodeID node (that is closest to and greater than the local node's NodeID) and L/2, and less than its nearest local node's NodeID of the node collection. NodeID and its nearest node is less than the local node's NodeID collection

(3). Neighbor collection.

Different from the numerical value approximation, the relationship between the neighbor collection and the set M, i.e., as far as the network in terms of measurement approaches is concerned. They are close to the current node. Thus, there is no routing itself, only in the maintenance of routing information in the local network.

3.2.2 Routing process

Routing in Pastry is divided into the following two steps:
(1) A node that is used to check the keyword k in its leaves is in the scope of the collection. Then k is located in the leaves set of a nearby node, therefore, the node will apply to transmit to the node numerically close to k in the most recent collection of nodes on the leaves. If it is found the node itself, the routing process is complete.
(2) If k did not fall into the scope of the collection node, then the routing table will use a longer distance to forward the request. In this case, the node n attempts to transmit the request to meet the following conditions of a node, that is, the nodes and K share a prefix longer than n itself and the K. In some cases, the routing table corresponds to table may be empty, or the best examples of the corresponding routing node unreachable. At this time the news will be forwarded to a node with the common prefix, but compared to the current node, the node will be more close to the keyword numerically. These kinds of nodes are located in a certain set of leaf nodes. Therefore, as long as the leaf nodes will not be set more than half of node failure at the same time, routing process can continue. From the above process in Pastry peer-to-peer network model, and compared with the previous step to the target nodes, we can see that every step of routing is making progress, so this process is convergent.
In Table 2 below, we show a router send a request of finding 103200 to 103210, as it is the closest leave from the keyword section of collection of nodes. Since the leaf node to hold the closest collection of nodes, so the keyword information is located on that node. Though the request of searching keyword 102022 is closer to 101203, but it is send to node 102303, this is because it shares the first 102 prefix (instead of the current node that shares the first 10 prefix). For keyword 103000, as there is no routing table that is longer than the current node to share the common prefix, and hence the current request will pass the node 103112, this node share 103 prefix, but its value is numerically more closer than the current node.

3.2.3 The self arrangement ability of pastry

The first mission of Pastry algorithm is to preserve the stability of Pastry system. Agreement is established from a stable system structure. In particular, the network has a property of high dithering, which means frequent node joining and withdrawing from the Pastry system, so preserving the stability of this system is much more important.

Assumed that the new joined node has a series number X (the number of node can be determined by Pl address or hashing SHA—1 Public key). Before X joining the pastry, we need to know a neighbor node A's location information. The process of adding X needs to initialize the data structure and notify the other nodes to join the system. Firstly, X should require that A road send a "to join" message, the keyword information is nodes number of X. The same with other information, this news will reach a node Z that has the closest node number with X. As a response, node A and node Z, and from A to Z on the path of all other nodes will take its own data structure to X. Then X use the information to initialize its own data structure, after the completion of initialization, X will notify other nodes that it has been joined the system. In order to handle concurrent node to join the system, Pastry uses a simple time-stamp mechanism. Simply speaking, when node A send its own data structure to node X, it is in the message a time stamp T_A attached. When the node X has completed its data structure initialization, a time stamp T_A is attached in its message sent to A.

In this way, node A will be able preserve its data structure information after checking in B. If the information is changed, then new information will be sent to the X to inform the re-initialization. Such a mechanism is proposed under the assumption that adding a small number of nodes. If at the same time a large number of nodes joining the system, the performance of this strategy need to be further studied.

There are situations that Pastry network node don't work or suddenly leave the system. When the adjacent nodes can not communicate with certain Pastry node, this nodded will be regarded as a failure node. If the L node fails in the leaf nodes set, then the current node will ask the largest or smallest node in the current leaf node set to sent its leaf set L (According to the failure node, if the number of failure node is larger than the current node, then use the node that has the largest node number, otherwise, use the node that has the smallest node number). If there is no node in set L, then current node will choose an alternative failure node. Before the replacement, we have to verify that the node is still in the system. If certain node in the router table does not work, then the current node will choose another node from the router table, and ask the new node to send its position term in the router. If there is no useful node in the corresponding row in the table, then the current node will choose another node from the next row, this process will continue until the current node failure can be an alternative node, or the current node has go through the whole router table.

3.2.4. The optimization of router

Pastry optimizes the router and position for a given keyword router. It tries to use the smallest jumpers to reach the objective node. And it can reduce the burden of each jumper, by using the internet local property.

(1). the length of the router.

Pastry routing mechanism actually divides the space into a space of 2n dimension, in which n is a multiple of 2b. Domain routing number from high to low, and therefore the left identifier space to be searched is reduced in every step. The intuitive result is that the average number of routing steps relates to the logarithm of the system size, and such kind of intuition is rational. Assume that all the routing information of nodes is correct, and there is no node break down. There are three cases in Pastry system. The first forwarding a request based on router table. In this case, the request will be forwarded to the node that a much more loner prefix is matched. Therefore, the number of nodes in each step will decrease at the speed of the factor of 2b.

Therefore, the request will reach the destination in $\log_{2^b}^{N}$ steps. The second case is to router a request by a leaf. The number of jumper will increase by one. In the third situation, the keyword will not covered by the leaves. And the routing table will not contain the matching prefix longer than the current node. As a result, the request will be forward to a node having the same length of prefix, with an additional router jumper is increased. For a middle size leaf collection of size $|L|=2*2$, the probability of such a situation is less than 0.6%, and therefore, the situation of additional jumper will almost never happens. As a result, the complexity of the router will preserve at O($\log_{2^b}^{N}$). The larger b is the faster router will be obtained. In the mean time, the additional management state will also be increased. Therefore, the typical value of b is 4, but Pastry will choose an appropriate compromise value for the typical application.

(2). The locality

By using the internet locality, Pastry does not only optimize jumpers, but also optimizes the cost of a single jumper. By making a criterion for positioning a router table, and allow to have a choice in nodes that has the matching prefix, the length of the single router will be minimized. Such kinds of methods may not generate the shortest router from end to end. But a more reasonable overall length will be generated.

In the initial identifier space, Pastry node will use a router table in the path from other node to the given node. The closeness of the new node n and the given node K implies that the first row of the table implying k also close to n. The nodes in the following rows of the path from k to n will be close to k, but not necessarily close to n. However, the distance from k to these nodes is longer than the distance from k to n. This is because the content in the router table have to choose from the set of small logarithm. Hence, their average distance from k to n will increase logarithm. Anther meaning of this fact is that the information path will increase with the router distance, they will become closer and closer to the objective ID.

4. Analysis of the algorithm

As the Pastry system routing algorithm has employ the maximum mask matching algorithm, so many known software and hardware frameworks can be used to achieve good efficiency. Compared with Chord, Pastry has introduced leaf node and neighbor nodes set, so when try to obtain the node information in the application layer, the searching speed of router can be accelerated. And the overhead of internet transport caused by router can be reduced. However, in the adaptive P2P network, it is quite difficult to accurately obtain the node sets of leaf and neighbor.

The main problem of DHT structure is that the maintenance of DHT mechanism is complex. The frequent joining in and exiting will increase the cost of maintain. If some Peer has problem, then the cost of maintenance will be quite expensive. Therefore, the structured P2P system des not adapt to the highly adaptive internet environment. Another problem that DHT face is that DHT only support the accurate keyword matching queries. The content/semantic and some other complex queries don't work. As the Pastry algorithm is done by the DHT approaches, so the half matching almost can not accomplished. We must provide the complete description of the searching resources. Such kinds of request seem unreasonable. Another problem is that the dimension of the structured P2P system is limited by its own algorithm, and therefore not suitable for large P2P system.

5. Concluding remarks

The mechanism of pasty system routing algorithm is analyzed in this paper. We point out the advantages and shortcomings of this algorithm compared with Chord algorithm. The Pastry algorithm of DHT is a hot topic, a series of important results have been obtained, and widely used in practical engineering. But further investigation is needed in semantic query routing algorithm, network stability and security remains for Pastry system routing algorithm.

References

Binzenhofer ,A, Staehle, D., &Henjes, R. (2005). Telecommunications Conference, GLOBECOM'05. IEEE Volume2, 28.

Huang, D. Y., & Li, Z. P. (2003). Active Distributed Peer-to-Peer Network Architecture, In: International Conference on Communication Technology Proceedings, 2003.

Rowstron, A. & Druschel, P. (2001). Pastry: A large-scale, persistent peer-to-peer storage utility", In HotOS VIII,School Elmau,Germany.

Rowstron, A. & Druschel. P. (2001). Pastry: Scalable, Distributed Object Location and Routing for Large-scale Peer-to-peer System, In IFIP/ACM international Conference on Distributed Systems Platforms (Middleware),Heidelberg,Germany,November 2001,springer, 329-350.

Stoica I. Morris R. Liben-Nowell D. Karger D R, Kaashoek MF, Dabek F, & Balakrishnan H. (2003). Networking, IEEE/ACM Transactions on, Volume 11, Issue 1.

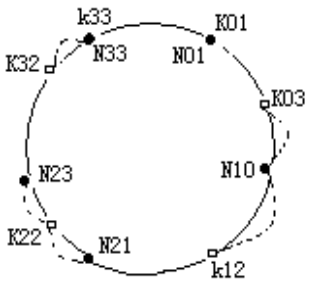

Figure 1. 4 bit Pastry identifier space

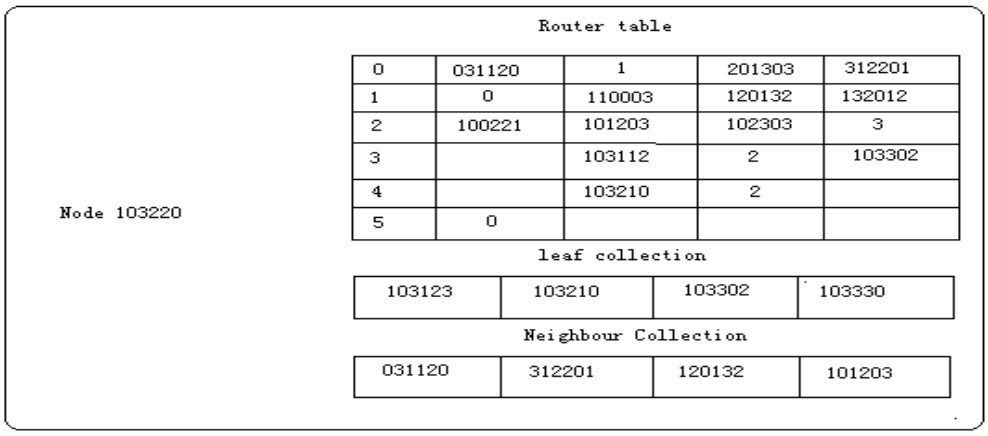

Figure 2. The state of 103220 Pastry node state in the case of 12 bit identifier space and 4 base

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 2, No. 4

November 2009

Research of Education Evaluation Information Mining Technology Based on Gray Clustering Analysis and Fuzzy Evaluation Method

Yang Liu

College of Computer and Automatization, Tianjin Polytechnic University

Tianjin, 300160, China

E-mail: lynn_liu11@yahoo.com.cn

Junle Yu

College of Application Technology, Tianjin Polytechnic University

Tianjin, 300020, China

E-mail: junleye@126.com

Fund Project: Tianjin Education Sciences, "Eleventh Five-Year Plan" issues, issues No.: G016

Abstract

This paper has surveyed the education evaluation method and technology both at home and abroad, and studied on the question of education evaluation information's mining and synthesis processing. In view of the evaluation index system of high and secondary vocational education, on the basis of gray clustering analysis method this paper had established gray clustering model and applied fuzzy evaluation method to solve the question of education evaluation information synthesis processing. It provided a vastitude future of education evaluation information's mining and synthesis processing.

Keywords: Gray Clustering, Education Evaluation, Fuzzy Evaluation, Data Mining

1. Introduction

Education evaluation is that according to particular education value and education object, using operable science artifice, make value judgment on education action, processing and result by systemic collecting, analyzing and arranging information data. The result of judgment will provide a basis for continuous self-improvement and education decision. The issue of education evaluation information processing and data mining is an important issue of education evaluation research The solution of this issue will not only promote education quality improvement but also provide valuable reference information for increasing university teaching innovation and implementing the quality education. At the abroad the theory and application of education evaluation had a quickly development from this century, but the method and theory of evaluation information's integrated processing and data mining was behindhand in development relative to education evaluation investigation. So far, it is three representative education evaluation patterns including object direction evaluation pattern, decision direction evaluation pattern, pluralistic evaluation pattern in the development of abroad education evaluation. The education evaluation pattern investigation noted very little in the method and theory of information integrated processing and data mining. Each education evaluation had a definite value orientation, and these value orientations or value judgment were dominated by the study and application of information integrated processing and data mining theory and method.

Education is a complex system, so the integrated evaluation of complex system was needed many index to weighing. A part of internal study on education evaluation and evaluation index system was based on analytical hierarchy process, and found more science evaluation index system. When more science evaluation index system was found, we must choose a more precisely computation method. Accordingly, the researcher who at abroad and home had pay more attention to how choose a more precisely computation method in the education evaluation processing. In 1982, the famous researcher named Julong Deng in our country had brought up the gray system theory. The research object of gray system theory was uncertainty system what is unknown part of information's small sample or poor information.

The gray clustering is the method that cluster several observation index or observation object into some definable classes by gray association matrix or gray number's whitely weight function. (Yannis Caloghirou. 1999)(Chen Z. 2001)(Qiu Jianrong, Zhang Xiaoping, Liu Hao, Wang Quanhai, Li Fan & Zeng Hancai. 2002)(Gu Zhaojun, He Xiaohui, Si Zhensheng & Fan Jingxin. 2007) A clustering can be considered as a set that observation was belonged same class. For the set what is constituted by all clustering object, we need not cluster by clustering index but also evaluate all evaluation object in a whole. For using data mining technology to point out the problem in the education evaluation and found the way to solve problem, this paper applies grey clustering analytic method as a basis and integration use fuzzy evaluation method(Sadaaki Miyamoto. 1990) to solve integration information processing in the education evaluation index system, thereby promote education evaluation's networking and informationization. At the same time according to high and secondary vocational education talented people training work level evaluation project, the paper use grey clustering method and fuzzy evaluation method to research and practice education evaluation in the evaluation project. The paper aspire after practicalness, pertinency and realistically in the content, materialize advanced and multiformity in the method, reach after science, justice and directional in the result. The research and solution of aforementioned problem will come about active effect and action for our country education evaluation technology.

2. Education evaluation information integration processing and data mining

The process of education evaluation data mining and information integration had three basic steps as follows:

2.1 Construct education evaluation index system hierarchy model and confirm index weight

After survey, education evaluation index system hierarchy's general model was confirmed by expert discussion again and again as figure 1 show:

We use AHP-GA method to confirm each index weight (also can be confirmed by experts researching). We assume that the weight vector what first index $B_1, B_2, ..., B_n$ relative to total object A is $b = (b_1, b_2, \ldots, b_n)$. The weight vector what layer C's weight relative to layer B's element B_n is $c_n = (c_1, c_2, \ldots, c_{cn})$. The cn means that the number of layer C's element relative to element B_n. And for the same reason, the weight vector what layer D's weight relative to layer C's element C_m is $d_m = (d_1, d_2, \ldots, d_{cm})$.

2.2 Construct grey clustering model

2.2.1 Construct evaluation value matrix

It is assumed that the number of evaluation person was p, namely $t = 1, 2, \ldots p$. The number of evaluation object was q, namely $s = 1, 2, \ldots q$. From above index system, we assume that the number of first index was n, the number of second index was m and the number of third index was k_m.

The evaluation person evaluates some evaluation object by third index's evaluation rank standard. We assume that the grade what the evaluation person t evaluates on evaluation object s by third index's evaluation rank standard was $d_{ijt}^{(s)}(i = 1, 2, \ldots, m; j = k_1, \ldots, k_i; t = 1, 2, \ldots, p; s = 1, 2, \ldots, q)$, so the evaluation object s's evaluation value matrix $D^{(s)}$ was:

$$D^{(s)} = \begin{bmatrix} d_{111}^{(s)} & d_{112}^{(s)} & \cdots & d_{11p}^{(s)} \\ & \cdots \ \cdots & & \\ d_{1k_1 1}^{(s)} & d_{1k_1 2}^{(s)} & \cdots & d_{1k_1 p}^{(s)} \\ & \cdots\cdots & & \\ d_{m11}^{(s)} & d_{m12}^{(s)} & \cdots & d_{m1p}^{(s)} \\ & \cdots\cdots & & \\ d_{mk_m 1}^{(s)} & d_{mk_m 2}^{(s)} & \cdots & d_{mk_m p}^{(s)} \end{bmatrix}$$

2.2.2 Confirm evaluation grey class

It should confirm the rank number of evaluation grey class, the grey number of grey class and grey class's whitely weight function by concretely education evaluation index system. It is assumed that the rank number of evaluation grey class was g, namely evaluation grey class was $e = 1, 2, \ldots, g$. The grey number is not a number, but it is a number set, or a space, is noted $\otimes$.

2.2.3 Compute grey class's evaluation coefficient $\eta_{ije}^{(s)}$.

For the third index, by the whitely weight function $f_e(d_{ijk}^{(s)})$ and the evaluation object s's value $d_{ijk}^{(s)}$, it computes grey evaluation coefficient $\eta_{ije}^{(s)}$ which valuation object s belongs to evaluation grey class e as follows

$$\eta_{ije}^{(s)} = \sum_{k=1}^{p} f_e(d_{ijk}^{(s)}) \tag{1}$$

2.2.4 Compute grey evaluation weight $r_{ije}^{(s)}$ and construct grey evaluation weight matrix.

The grey evaluation weight $r_{ije}^{(s)}$ what evaluation person maintained grey e on evaluation object s by third evaluation indexes was:

$$r_{ije}^{(s)} = \eta_{ije}^{(s)} \Big/ \sum_{e=1}^{g} \eta_{ije}^{(s)} \tag{2}$$

After colligating all grey class's evaluation object s toward third evaluation indexes, the grey evaluation weight vector $r_{ij}^{(s)}$ is:

$$r_{ij}^{(s)} = (r_{ij1}^{(s)}, r_{ij2}^{(s)}, \cdots, r_{ijg}^{(s)}) \tag{3}$$

By colligating grey evaluation weight $r_{ije}^{(s)}$ in the all of second index C_i's third index, it can be found that grey evaluation weight matrix $R_i^{(s)}$ of evaluation object s's second index C_i toward each evaluation grey class is:

$$R_i^{(s)} = \begin{bmatrix} r_{i1}^{(s)} \\ r_{i2}^{(s)} \\ \vdots \\ r_{ik_i}^{(s)} \end{bmatrix} = \begin{bmatrix} r_{i11}^{(s)} & r_{i12}^{(s)} & \cdots & r_{i1g}^{(s)} \\ r_{i21}^{(s)} & r_{i22}^{(s)} & \cdots & r_{i2g}^{(s)} \\ \vdots & \vdots & \vdots & \vdots \\ r_{ik_i1}^{(s)} & r_{ik_i2}^{(s)} & \cdots & r_{ik_ig}^{(s)} \end{bmatrix}$$

2.2.5 Compute grey integration evaluation vector $TT_i^{(s)}$ of second index C_i.

For the evaluation object s, according to $d_i = (d_1, d_2, \ldots, d_{ci})$ that layer D's relatively weight vector corresponding layer C's element C_i and grey evaluation weight matrix of second index C_i toward each evaluation grey class, it can be evaluated integration, and then it found that grey integration evaluation vector $TT_i^{(s)}$ of evaluation object s's second index C_i as follows:

$$TT_i^{(s)} = d_i \cdot R_i^{(s)} = (tt_{i1}^{(s)}, tt_{i2}^{(s)}, \cdots, tt_{ig}^{(s)}) \tag{4}$$

2.2.6 Compute grey integration evaluation vector $LT_i^{(s)}$ of first index B_n.

For the evaluation object s, by colligating grey integration evaluation vector $TT_i^{(s)}$ in the each second index C_i, it can be found that grey evaluation weight matrix $L_i^{(s)}$ of evaluation object s's first index B_n toward each evaluation grey class is:

$$L_i^{(s)} = \begin{bmatrix} TT_1^{(s)} \\ TT_2^{(s)} \\ \vdots \\ TT_m^{(s)} \end{bmatrix} = \begin{bmatrix} tt_{11}^{(s)} & tt_{12}^{(s)} & \cdots & tt_{1g}^{(s)} \\ tt_{21}^{(s)} & tt_{22}^{(s)} & \cdots & tt_{2g}^{(s)} \\ \vdots & \vdots & \vdots & \vdots \\ tt_{m1}^{(s)} & tt_{m2}^{(s)} & \cdots & tt_{mg}^{(s)} \end{bmatrix}$$

According to $c_n = (c_1, c_2, \ldots, c_{cn})$ that layer C's relatively weight vector corresponding layer B's element B_n and grey evaluation weight matrix of first index B_n toward each evaluation grey class, it can be evaluated integration, and then it found that grey integration evaluation vector $LT_i^{(s)}$ of evaluation object's first index B_n as follows:

$$LT_i^{(s)} = c_n \cdot L_i^{(s)} = (lt_{i1}^{(s)}, lt_{i2}^{(s)}, \cdots, lt_{ig}^{(s)}) \tag{5}$$

2.2.7 Compute clustering result.

For the evaluation object s, by colligating grey integration evaluation vector $LT_i^{(s)}$ in the each first index B_i $(i = 1, 2, \ldots, n)$, it can be found that grey evaluation weight matrix $Z^{(s)}$ of evaluation object s toward each evaluation grey class is:

$$Z^{(s)}=\begin{bmatrix} LT_1^{(s)} \\ LT_2^{(s)} \\ \vdots \\ LT_m^{(s)} \end{bmatrix}=\begin{bmatrix} lt_{11}^{(s)} & lt_{12}^{(s)} & \cdots & lt_{1g}^{(s)} \\ lt_{21}^{(s)} & lt_{22}^{(s)} & \cdots & lt_{2g}^{(s)} \\ \vdots & \vdots & \vdots & \vdots \\ lt_{m1}^{(s)} & lt_{m2}^{(s)} & \cdots & lt_{mg}^{(s)} \end{bmatrix}$$

According to the weight vector $b=(b_1,b_2,\ldots,b_n)$ of first index B_i $(i=1,2,\ldots,n)$ and grey evaluation weight matrix $Z^{(s)}$ of evaluation object s, it can be clustered integration, and then it found that clustering result $X^{(s)}$ of evaluation object as follows:

$$X^{(s)}=b\cdot Z^{(s)}=(x_1^{(s)},x_2^{(s)},\cdots,x_g^{(s)}) \tag{6}$$

2.3 Apply fuzzy evaluation to finding evaluation result

It uses each grey class's threshold as rank value to compute the integration evaluation value of each evaluation object. For example, the grey class one's threshold is d_1, the grey class two's threshold is d_2,…, the grey class g's threshold is d_g. So each grey class rank value vector is $F=(d_1,d_2,\cdots,d_g)$. The integration evaluation value $G^{(s)}$ of evaluation object s is:

$$G^{(s)}=X^{(s)}\cdot F^T \tag{7}$$

It can be made an order by the value of $G^{(s)}$ after computing each evaluation object's integration evaluation value $G^{(s)}$.

3. Demonstration test

We use the evaluation process of high and secondary vocational education talented people training work level evaluation index system as example to explain the process of education evaluation information integration processing and data mining. According to the model of figure 1, firstly we have to confirm evaluation index's weight. Secondly, it evaluated on school A by five experts (data in table 1), then it found evaluation value matrix as follows:

$$D=\begin{pmatrix} 8 & 7.5 & 8.5 & 8 & 7 \\ 7 & 6.5 & 7 & 6 & 7 \\ 6 & 6.5 & 7 & 6 & 6 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 8 & 8.5 & 9 & 8 & 7.5 \\ 7.5 & 8 & 7 & 6 & 7 \end{pmatrix}$$

According to grade standard of high and secondary vocational education talented people training work level evaluation project, it can be confirmed that the evaluation grey number is $e=5$, the whitely weight function is $f_1(x),f_2(x),f_3(x),f_4(x),f_5(x)$ as follows:

$$f_1(x)=\begin{cases} 1 & x\in[9,+\infty]; \\ \frac{1}{9}x & x\in[0,9] \end{cases} \qquad f_2(x)=\begin{cases} 1 & x\in[8,9]; \\ \frac{1}{8}x & x\in[0,8]\cup[9,+\infty] \end{cases}$$

$$f_3(x)=\begin{cases} 1 & x\in[7,8]; \\ \frac{1}{7}x & x\in[0,7]\cup[8,+\infty] \end{cases} \qquad f_4(x)=\begin{cases} 1 & x\in[6,7]; \\ \frac{1}{6}x & x\in[0,6]\cup[7,+\infty] \end{cases}$$

$$f_5(x)=\begin{cases} 1 & x\in[0,6]; \\ \frac{1}{6}x & x\in[6,+\infty] \end{cases}$$

It can be found that school A's all second index toward each evaluation grey class $R_i^{(s)}$ is (use first second index as example):

$$R_1=\begin{bmatrix} r_{11} \\ r_{12} \\ r_{13} \end{bmatrix}=\begin{bmatrix} 0.587 & 0.205 & 0.102 & 0.100 & 0.006 \\ 0.532 & 0.247 & 0.112 & 0.102 & 0.007 \\ 0.514 & 0.234 & 0.152 & 0.100 & 0 \end{bmatrix}$$

Then, it can be found that grey evaluation weight matrix $L_i^{(s)}$ of first index's each evaluation grey class is (use first of first index as example):

$$L_1=\begin{bmatrix} TT_1 \\ TT_2 \end{bmatrix}=\begin{bmatrix} 0.524 & 0.245 & 0.143 & 0.080 & 0.008 \\ 0.562 & 0.217 & 0.115 & 0.102 & 0.004 \end{bmatrix}$$

Final, school A's clustering result is:

$X^{(A)} = (0.5521, 0.2548, 0.1821, 0.007, 0.004)$

Namely school A was belonged to first clustering (general comment above nine points), then it can be used fuzzy evaluation method to compute concretely general comment value as follows:

$G^{(A)} = 9.3751$

According to above computing, it can be found that school A's evaluation result as table1 shows:

4. Conclusion

To sum up, in the practice application grey clustering method was relative agility. We had colligated fuzzy evaluation method and grey clustering method to clustering analyze and fuzzy evaluate on experts data; take integration analysis and advisement on evaluation index system's data. This method was more exactitude, more applied and more abundant what information shows compare with weighted averages method and other computation method. Foreign and Chinese had worked hard on the theory and application of grey clustering method. So we try to use this method to integrate process on education evaluation information in this paper, and the result was reasonable. The research result aspires after practicalness, pertinency and realistically in the content, materialize advanced and multiformity in the method, reach after science, justice and directional in the result, provide more strict and scientific method for the research of high and secondary vocational education evaluation information integration processing. At the same time, based on the research result building education evaluation database, evaluation model database and education evaluation information integration processing system on the Web was feasible and necessary. The research result was established stability basic for the development of our country's education evaluation technology.

References

Chen Z. (2001). An application of grey clustering method in the sporting clothing style evaluation. *Journal of Systems Engineering and Electronics*, v 12, n 2 (2001)19-22.

Gu Zhaojun, He Xiaohui, Si Zhensheng & Fan Jingxin. (2007). Research on spam filtering based on grey clustering. *Journal of Computational Information Systems*, v 3, n 4 (2007)1713-1718.

Qiu Jianrong, Zhang Xiaoping, Liu Hao, Wang Quanhai, Li Fan & Zeng Hancai. (2002). Grey clustering prediction for slagging potential of coal blends combustion. *Combustion Science and Technology*, v 174, n 3 (2002)51-70.

Sadaaki Miyamoto. (1990). *Fuzzy sets in Information retrieval and Cluster Analysis* (1st ed). Kluwer Academic Publishers.

Yannis Caloghirou. (1999). Mutivariate analysis for assessment of corporate performance. the case of Greece. *Operational Tools in the Management of financial risks.*

Table 1. School A's experts grading and General comment

	E 1	E 2	E 3	E 4	E 5
School orientation and develop programming	8	7.5	8.5	8	7
Education thinking concept	7	6.5	7	6	7
Teaching center position	6	6.5	7	6	6
Study and teaching	8	7.5	9	9	8
proportion between student and teacher	8	8	8.5	8	7.5
Non-plurality teacher proportion	9	8.5	9	8	8.5
plurality teacher's quantity and structure	8	8.5	9	8	7.5
Quality	8	8.5	9	7.5	8

Construction and development	9	8.5	9	8.5	8
Teaching administration house	7	8	8.5	9	8
Teaching instrument equipment	8	9	8.5	8	7.5
Library and campus net	7	6	6.5	8	7.5
Athletic sports establishment	9	8.5	8	7.5	8.5
Practice condition in school	8	8.5	8.5	9	7.5
Practice base out of school	7	8	8.5	7.5	7.5
Profession skill appraisal	6	8	7	6.5	7
Outlay pledge complexion	8.5	9	9.5	8	8.5
Teaching outlay proportion	6.5	7	7.5	6	7
Major setting	7	8.5	8	7.5	9
Teaching plan	8	8.5	9	7.5	8.5
Major teaching innovation	9.5	9	8.5	9	9.5
Teaching content and course innovation	8.5	9	8.5	8	9
Teaching material construction	7	8	8.5	7	9
Teaching method	7.5	8.5	8	7.5	8
Practice training system	8	8.5	9	7.5	8
Profession ability checking	8.5	8	8.5	9	8
Work state and impact on all around education	7.5	8	8.5	9	8
Teaching manage and student manage	8	9	8	9	9.5
Teaching bylaw's construction and perform	8.5	8	9	8.5	8
Quality standard of major teaching tache	7	7	7	8	7.5
Teaching quality monitor and student quality research	7.5	8	8.5	7	7.5
Profession ability	6	8	6.5	6	7
Necessary knowledge	8.5	8.5	8.5	9	7
Basic diathesis	7	7.5	6	8	7
Register rate and employment rate	8	8.5	9	8	7.5
Graduate evaluation	7.5	8	7	6	7
General comment	9.3751				

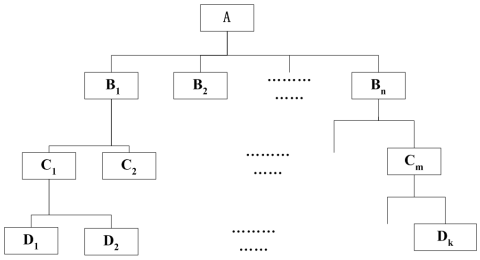

Figure 1. Education evaluation system hierarchy diagram

Computer and Information Science
www.ccsenet.org/journal.html
Vol. 2, No. 4
November 2009

The Application of the Real-time Temperature Monitoring System for Electric Transmission Lines

Ruicheng Li
Shandong Youth Administrant Cadre College, Jinan 250103, China
Tel: 86-531-5899-7405 E-mail: LRC04@163.com

Abstract

In the article, the influencing factors of the high-voltage transmission line running were analyzed, the important meanings of the real-time online monitoring of the line contact point aging for guaranteeing the safe and stable running of the lines were pointed out, the theoretical reference and the project selection of the high-voltage transmission line temperature monitoring were discussed, and the structure composing, working principle and running mode of the project proposed by the author were mainly introduced. The system designed in the article could prejudge the faults of the contact points and implement the state overhaul.

Keywords: High-voltage electric transmission line, Real-time temperature monitoring, Contact point aging

1. Backgrounds analysis

Overhead high-voltage transmission lines are the arteries of the electric power system, and their running states directly decide the safety and benefits of the electric power system. The contact point breaking is one of usual fault existing on the overhead electric transmission lines.

The mechanical connection parts of the connection electric power fittings such as double-line yoke plate (parallel cable clamp), strain clamp and connecting pipe in the high-voltage transmission lines often have many thermal defects because of oxidation corrosion, loose connection or bad fixing quality, and when the power transmission lines run, the temperature of these parts will rise, and the aging of these parts will be pricked up, and the contact resistances will further increase, and finally the lines will be broken suddenly. So the aging of the contact points of the lines should be monitoring online and warned at any moment to find out the thermal defects in the electric power fittings, implement the state overhaul, predict and prevent the occurrences of the power breaking accidents, and ensure the safe and stable running of the power transmission lines (He, 1987).

According to the experiences and experts' analysis, there are three reasons to reduce the overheating of the line contact points, i.e. the oxidation corrosion, loose connection and bad fixing quality. At present, there are two methods to judge the thermal defect. The first method is the warning temperature rise method which takes the temperature rise comparing with the temperature of the environment as the reference, and if the temperature rise exceeds the warning temperature rise in the standard warning temperature rise table, the thermal defect occurs. For the method, the environment temperature surrounding the lines is difficult to be measured, and the solar radiation and the influences of electric power fitting can not be considered. The another method is the relative temperature rise method which judges the thermal defect by analyzing the change relation between the relative temperature difference and the contact resistance, but in practice, the contact resistance is very hard to be measured (Zhang, 1991 & Cen, 1996).

Because of the deficiencies of above two methods, a new relative temperature difference method occurs, which can simultaneously measure the temperatures of the heating points, the neighboring lines and the environment, and compute the relative difference according to the following formula.

$$\delta = (T1-T2)/(T1-T0)*100\%$$

Where, T1 is the temperature of heating point, T2 is the temperature of the corresponding point and T0 is the temperature of the environment. The judgment references of the guide equipment defect include the common defect of "≥35%", the major defect of "≥80%" and the equal emergence defect of "≥95%".

This method can be used to eliminate the additive temperature rise induced by the solar radiation, and reduce the errors because of inexact parameters such as the detecting range, environment temperature, humidity and wind speed. This

method has good effect to judge the thermal defect in the practice.

The common fault exists in above all methods, i.e. the manual work needs to be used in the spot, but the electric transmission lines are distributed in the very wide regions, so these methods are meaningless for the lines checking with large region. The real-time and online function actualized in the system could solve this difficult fault.

2. Design of the temperature monitoring project

The locale sketch map measuring the contact point temperature of the high-voltage overhead lines is seen in Figure 1. The contact points of the lines are generally located near the pole tower, and the lines are hung on the pole tower by two clusters of insulator, and each phase of line have 2~10 contact points.

First, the temperature can be measured by personnel at the locale. The personnel can adopt two sorts of method. The first one is the remote viewing temperature coating method, i.e. several temperature coats with different colors are adsorbed on the line contact points in advance, and these temperature coats will melt under different temperatures, and the melting state of each coat can be observed by the telescope to judge the approximate temperature range of the contact point. Another method is to utilize the remote infrared video camera which can only exactly observe the temperature only in the cloudy weather, not in the rain day, and the errors and the costs are large. Above two methods all have their fixed deficiencies, i.e. they are largely influenced by the weather, and the measurement can not be implemented at any moment, and most lines are distributed in the wider regions, and the measurement will waste large numbers of manpower and materials.

Second, the method of real-time online monitoring can be adopted, in which there are three modes including infrared mode, optical fiber mode and wireless to be selected mode (Chen, 2003).

The infrared mode is to utilize the infrared temperature sensor installed on the remote end to measure the temperature of the point, and this mode can not be jammed by the electromagnetic field, and it needs not the contact with the personnel. But for the high-voltage electric transmission lines, because the measured object is too small and the measurement distance can not be too close because of the installation limitation, so the distance coefficient of the sensor D: S (the ratio between the distance and the diameter of the measured object) is required to be quite big, but there is the proper product to fulfill these requirements. A two-color thermometric indicator can measure the temperature of the remote small object, but it can only aim at the measurement object above 500℃. In addition, the infrared sensor can only be installed on the pole tower by the side of the lines, and the obstacles can not be located among the measured points, and the telescope needs to exactly aim at the object, which is difficult to be implemented in practice.

The optical fiber mode is to directly install the optical fiber sensor on the contact points of the lines, and transfer the data to the lower computer by the optical fiber. The optical fiber has the high insulated property, and it can not be jammed by the electromagnetic field, and it can be applied in the indoor high-voltage, but in the field, the rains and dews will reduce the insulated performances if they attach the surface of the optical fiber and induce the safety fault. One remedial method is to connect a cluster of insulators among the optical fibers, which will be too heavy, and fixed difficultly, and to monitor the contact points of the lines, tens of points should be measured by the side of each pole power. In above two methods, one point needs one sensor, and the costs of two methods are too high (Jia, 1998, P.61-66 & Zhang, 2004).

3. System design

Through comprehensively considering, a wireless temperature measurement method is designed, i.e. a micro control model is installed on each phase line, and the model is connected with many temperature collecting detector to collect the temperatures of many contact points and lines on the phase line, and the model transfers the collected temperature data to the equipment installed on the pole tower. This method will never influence the safety performance of the electric transmission lines, and through corresponding technical measure, the method can solve the problems that the wireless communication is jammed by the electromagnetic filed near the lines, and the power supply of the temperature measurement model and the temperature collection precision are easily be influenced by the high-voltage electric filed. In addition, the cost of this method is very low.

In the system, the networking mode that one monitoring center host machine matches with many outdoor slave machines (<100 computers) is adopted, and the wireless data communication is used by the GSM network. The system networking is seen in Figure 2.

Because the GSM network is used for the communication and the distances between the host machine and slave machines are not limited, so the system can be used in the regions that the GSM network could cover. This method can solve the problem of the networking distribution of lines, and the monitoring center can be moved at any moment without outdoor antennas with huge cubage and difficult installation, which makes the system more flexible.

The system can measure many parameters such as the contact point temperature of the high-voltage lines, the temperature of the lines, the temperature of the environment, the humidity, the wind direction and the wind speed, and

one outdoor slave machine installed on the pole tower can monitor 1~32 line contact points.

3.1 System composing

The system includes the outdoor slave machine, the center host machine and the system software.

The outdoor slave machines are installed on the pole tower of the field high-voltage transmission lines, and they are used to collect the data of various parameters which are transferred to the monitoring center after being processed, and they receive various orders from the center at the same time to implement the operations such as point measuring, modifying threshold setting, modifying time interval, and collating the clock. The powers of the outdoor machines are supplied by the solar battery with the intelligent charge management system which needs not maintenances for many years, and the life of the battery is 3~6 years. The micro-power-consumption UHF wireless receive chip is adopted between the outdoor slave machines and the temperature collection point to exchange the data and ensure the safety.

The monitoring center host machine is composed by the wireless receive equipment (GSM Modem) and the monitoring computer. The monitoring computer includes one industrial control computer server and many client computers which can read data on the server through the networking. The functions of the server and the client machines can be completed by one machine.

The system software is divided into three modules including the operation processing platform, the service control center and the database system, and these three modules can be respectively installed in different PCs or be installed in one PC. The software system could analyze and process the received contact point temperature and other data, and compute the temperature rise of the contact point according to the mathematical model, and offer the maximum temperature and the average temperature in certain period and the comparisons with the temperatures under other weather conditions, which could be referred for the monitoring personnel. The software includes the warning system which can alarm the monitoring personnel by the sound, light, or the short message. In addition, the software uses the advanced MapInfo geographic information system which can clearly display the actual geographic positions of various monitoring points, and amplify or shrink the map infinitely, and realize the intuitional human-machine interface.

3.2 Working mode

The working mode of the system is divided into the normal working, the warning and actively uploading, and the host machine transmitting orders.

When the system works normally, the outdoor slave machines check the parameters such as the temperatures and humidity degrees of various points, the wind direction and the wind speed periodically, compute, process and pack the data and transmit the data to the monitoring host machine according to certain time interval by the form of short message, and the system software of the center will process, store and displace the data. The time interval here can be set and modify at any moment by the management personnel through the system software.

The warning and actively uploading mode is that when the outdoor slave machine judges and checks that the temperature data exceed the threshold value set up in advance, it actively upload the data immediately.

The host machine transmitting orders mode means that the host machine can make various orders to the outdoor salve machines, and these orders include point measuring, collating clock, modifying the set of threshold, resuming the default setting and modifying the time interval that the salve machine transmits the data. And the point measuring requires that the slave machine measures the data and returns the measurement values at once, and the collating clock means the slave machine collating the clock to make all equipments in the system work at same clock system. The GSM module of the outdoor machine is always in the receiving state, and when it receives the orders of the host machine, it should respond immediately, complete the appointed operation and return corresponding information.

The contact point temperature measurement system is the difficulty in the design. The temperature collector and the wireless receiver must be installed on the high-voltage lines, but they can not use the power supply of the high-voltage lines, and they will be jammed by the stronger electromagnetic filed and the high-frequency harmonics, and once they are installed, they are difficult to be maintained, which bring more research tasks for this circuit. In the design, the special-made solar batter is adopted to supply the power for the module, and the special protection equipments for overcharge, over-discharge, and discharge times limitation are designed to protect the circuit. The power supply of the whole module is controlled by the clock chip, and the special control protocol is adopted to wirelessly receive the data to save the electric power and transfer the data reliably at any moment. The temperature collection doesn't use traditional analog signals, but the high-integrated digital temperature sensor which can not only transfer the digital signals but enhance the level of the signal to prevent being submerged by the strong noises of the high-voltage lines. All circuits in the system are screened specially, and the wireless communication uses UHF frequency and adopts multi-levels anti-jamming processing. In addition, to achieve the special application, all chips in the module design are the most advanced, high integrated, ultra low power consumption, anti-jamming and ultra small volume, which makes the whole collection system more artful and easily to be installed, and work stably without maintenances.

4. Conclusions

Through the usage of the system, the temperature change rule of the high-voltage overhead line contact point can be simply grasped, which can be used to prejudge the fault of the contact point and realize the state overhaul, and the relative reliability that the connection electric fittings are processed by the exploding press technology at present is proved. This system with friendly human-computer interface and powerful communication function can be applied in various ultra high-voltage transmission line protections, and it can not only offer intuitional judgment references for the management department and the first-hand data for the technical improvements in relative industries, but save large numbers of human powers and materials and produce large social and economic benefits.

References

Cen, Amao. (1996). *Overhead Transmission Lines Construction Technologies.* Ningbo: Ningbo Publishing House.

Chen, Jiabin. (2003). *Electrical Equipment Malfunction Detection and Diagnosis Method and Examples.* Beijing: China Water Power Press.

He, Jiali & Ge, Yaozhong. (1987). *Ultra High Voltage Transmission Line Malfunction Analysis and Relay Protection.* Beijing: Science Press. No.24.

Jia, Junguo, Fan, Yunpeng & Li, Jing. (1998). The Transmission Line Malfunction Ranging Technology and Application Based on Current Traveling Wave. *Power System Technology.* No.22(8). P.61-66.

Zhang, Diansheng. (1991). *Design Manual of Electric Engineering High Voltage Transmission Lines.* Beijing: China Electric Power Press.

Zhang, Shuqi. (2004). About the Breakdown of the Line and Measures of Electric Power Supply Reliability. *Northwest China Electric Power.* No. 6.

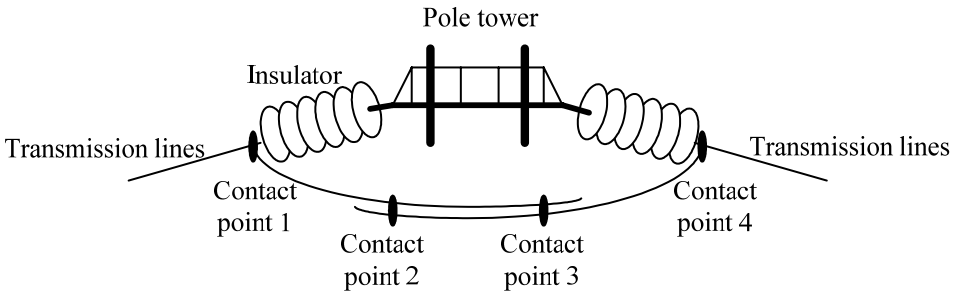

Figure 1. Sketch Map of High Voltage Transmission Line Contact Point Positions

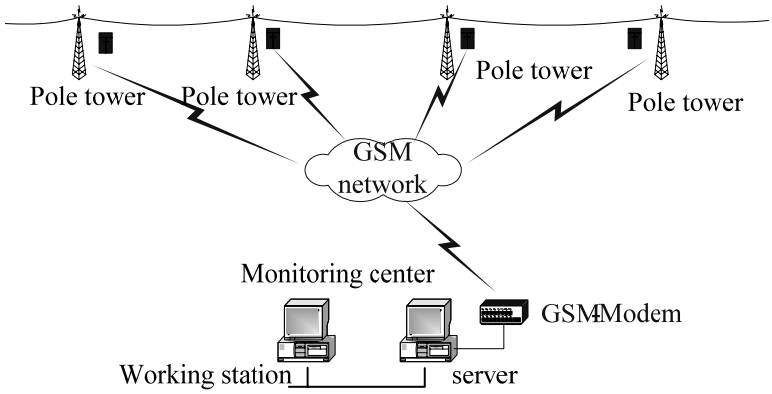

Figure 2. Sketch Map of System Networking

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 2, No. 4

November 2009

Two-Dimensional Heteroscedastic Discriminant Analysis for Facial Gender Classification

Junying Gan

School of Information, Wuyi University

Jiangmen, Guangdong 529020, China

E-mail: jygan@wyu.cn, junyinggan@163.com

Sibin He

School of Information, Wuyi University

Jiangmen, Guangdong 529020 China

E-mail: hesibin123@163.com

This work is supported by NSF of Guangdong Province, P.R.C. (No.07010869), by the fund of National Laboratory on Machine Perception (No.0505), Peking University, and State Key Lab of CAD &CG (No.A0703), Zhejiang University.

Abstract

In this paper, a novel discriminant analysis named two-dimensional Heteroscedastic Discriminant Analysis (2DHDA) is presented, and used for gender classification. In 2DHDA, equal within-class covariance constraint is removed. Firstly, the criterion of 2DHDA is defined according to that of 2DLDA. Secondly, the criterion of 2DHDA, log and rearranging terms are taken, and then the optimal projection matrix is solved by gradient descent algorithm. Thirdly, face images are projected onto the optimal projection matrix, thus the 2DHDA features are extracted. Finally, Nearest Neighbor classifier is selected to perform gender classification. Experimental results show that higher recognition rate is obtained by way of 2DHDA compared with 2DLDA and HDA.

Keywords: Gender classification, Two-dimensional teteroscedastic discriminant analysis, Two-dimensional linear discriminant analysis

1. Introduction

Gender classification using face images is a challenging work due to the similarity between male and female face images. Thus, discriminant feature extraction is a key step to improve recognition rate. Linear Discriminant analysis (LDA) is a well-known approach for feature extraction and dimensional reduction. However, it often encounters the Small Sample Size problem (S3 problem) when the number of samples is less than the dimensionality of samples. Then, two-dimensional Linear Discriminant analysis (2DLDA) is proposed, in which discriminant features are extracted directly from 2-D images without a vectorization procedure, the computation cost is reduced and the S3 problem is overcome. However, in both of LDA and 2DLDA, it is assumed that the covariance matrices are equal for all sample classes. Thus, when the within-class covariance of each sample class is significantly unequal, optimal performances can not be gained by LDA and 2DLDA.

Heteroscedastic Discriminant analysis (HDA) is extended from LDA, in which equal within-class covariance constraint is removed. HDA can be viewed as a constrained Maximum likelihood (ML) projection, the constraint is given by the maximization of the projected between-class covariance volume and each class a single full covariance Gaussian model is satisfied. HDA is widely used in speech recognition and recognition rate is greatly increased than that of LDA. But in 1D-based approaches, the transformation matrix is difficult to calculate due to high dimensionality and extreme sparseness of the data. In this paper, based on 2DLDA and HDA, two-dimensional Heteroscedastic Discriminant analysis (2DHDA) is presented and used for gender classification. Firstly, the criterion of 2DHDA is defined, and log and rearranging terms are taken, then optimal projection matrix is solved by gradient descent algorithm. Secondly, face images are projected onto the optimal projection matrix, thus the discrimination features of face images are extracted.

Finally, Nearest Neighbor classifier is selected to perform gender classification**.** Experimental results show the validity of 2DHDA method.

2. Presented Approach

Suppose there are C sample classes, represented by $\boldsymbol{A}^1, \boldsymbol{A}^2, \boldsymbol{A}^3, \mathrm{L}\ , \boldsymbol{A}^c$ respectively. The total number of samples is N and each class includes n samples, that is $nc = N$. $\boldsymbol{A}_j^i \in \mathbf{R}^{m \times l}$ denotes the j th $(j = 1,2,3,\mathrm{L}\ ,n)$ sample which belongs to the i th $(i = 1,2,3,\mathrm{L}\ ,C)$ class. Thus, the mean of the i th sample class is $\overline{\boldsymbol{A}^i} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{A}_j^i$, and the global mean of all samples is $\overline{\boldsymbol{A}} = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{n} \boldsymbol{A}_j^i$.

2.1 2DLDA Approach

2DLDA's criterion is defined as

$$J(\boldsymbol{\theta}_{2\mathrm{DLDA}}) = \frac{\left|\boldsymbol{\theta}_{2\mathrm{DLDA}}^{\mathrm{T}} \boldsymbol{S}_b \boldsymbol{\theta}_{2\mathrm{DLDA}}\right|}{\left|\boldsymbol{\theta}_{2\mathrm{DLDA}}^{\mathrm{T}} \boldsymbol{S}_w \boldsymbol{\theta}_{2\mathrm{DLDA}}\right|} \tag{1}$$

where $\boldsymbol{S}_w$ is called within-class covariance matrix and $\boldsymbol{S}_b$ is called between-class covariance matrix of training samples, expressed respectively as

$$\boldsymbol{S}_w = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{n}\left(\boldsymbol{A}_j^i - \overline{\boldsymbol{A}^i}\right)^{\mathrm{T}}\left(\boldsymbol{A}_j^i - \overline{\boldsymbol{A}^i}\right) \tag{2}$$

$$\boldsymbol{S}_b = \frac{1}{N}\sum_{i=1}^{c} n\left(\overline{\boldsymbol{A}^i} - \overline{\boldsymbol{A}}\right)^{\mathrm{T}}\left(\overline{\boldsymbol{A}^i} - \overline{\boldsymbol{A}}\right) \tag{3}$$

Transformation matrix $\boldsymbol{\theta}_{2\mathrm{DLDA}}$ is calculated by the solution of the eigenvalue and eigenvector problem of $\boldsymbol{S}_b \boldsymbol{S}_w^{-1}$.

2.2 2DHDA Approach

2DHDA is the heteroscedastic extension of 2DLDA. In 2DHDA, equal within-class covariance constraint is removed and the criterion is defined which maximizes the class discrimination in the projected subspace. The criterion of 2DHDA is defined as

$$J\left(\boldsymbol{\theta}_{2\mathrm{DHDA}}\right) = \prod_{i=1}^{C}\left(\frac{\left|\boldsymbol{\theta}_{2\mathrm{DHDA}}^{\mathrm{T}} \boldsymbol{S}_b \boldsymbol{\theta}_{2\mathrm{DHDA}}\right|}{\left|\boldsymbol{\theta}_{2\mathrm{DHDA}}^{\mathrm{T}} \boldsymbol{W}_i \boldsymbol{\theta}_{2\mathrm{DHDA}}\right|}\right)^n = \frac{\left|\boldsymbol{\theta}_{2\mathrm{DHDA}}^{\mathrm{T}} \boldsymbol{S}_b \boldsymbol{\theta}_{2\mathrm{DHDA}}\right|^N}{\prod_{i=1}^{C}\left|\boldsymbol{\theta}_{2\mathrm{DHDA}}^{\mathrm{T}} \boldsymbol{W}_i \boldsymbol{\theta}_{2\mathrm{DHDA}}\right|^n} \tag{4}$$

where $\boldsymbol{W}_i = \frac{1}{n}\sum_{j=1}^{n}\left(\boldsymbol{A}_j^i - \overline{\boldsymbol{A}^i}\right)^{\mathrm{T}}\left(\boldsymbol{A}_j^i - \overline{\boldsymbol{A}^i}\right)$ denotes the covariance matrix of the i th sample class. Thus, $\boldsymbol{S}_w = \frac{1}{C}\sum_{i=1}^{c} \boldsymbol{W}_i$. According to equation (1) and (4), if covariance matrix $\boldsymbol{W}_i$ of all sample classes is assumed equal, then $J\left(\boldsymbol{\theta}_{2\mathrm{DHDA}}\right) = \left(J(\boldsymbol{\theta}_{2\mathrm{DLDA}})\right)^N$. However, $\boldsymbol{\theta}_{2\mathrm{DLDA}}$ is invariant to scale transformation of $J(\boldsymbol{\theta}_{2\mathrm{DLDA}})$, then $\boldsymbol{\theta}_{2\mathrm{DHDA}} = \boldsymbol{\theta}_{2\mathrm{DLDA}}$ is satisfied and 2DHDA is become 2DLDA. By taking log and rearranging terms, we get

$$H\left(\boldsymbol{\theta}_{2\mathrm{DHDA}}\right) \square\ N\log\left|\boldsymbol{\theta}_{2\mathrm{DHDA}}^{\mathrm{T}} \boldsymbol{S}_b \boldsymbol{\theta}_{2\mathrm{DHDA}}\right| - \sum_{i=1}^{C} n\log\left|\boldsymbol{\theta}_{2\mathrm{DHDA}}^{\mathrm{T}} \boldsymbol{W}_i \boldsymbol{\theta}_{2\mathrm{DHDA}}\right| \tag{5}$$

H has two useful properties of invariance[5]. For every nonsingular matrix $\boldsymbol{\varphi} \in \mathbf{R}^{l \times l}$, $H(\boldsymbol{\varphi}\boldsymbol{\theta}_{2\mathrm{DHDA}}) = H(\boldsymbol{\theta}_{2\mathrm{DHDA}})$. This means that subsequent feature space transformations of the range of c will not affect the value of the criterion. The second is that the criterion is invariant to row or column scalings of $\boldsymbol{\theta}_{2\mathrm{DHDA}}$ or eigenvalue scalings of $\boldsymbol{\theta}_{2\mathrm{DHDA}}\boldsymbol{\theta}_{2\mathrm{DHDA}}^{\mathrm{T}}$. Using matrix differentiation, the derivative of H is given by

$$\frac{\partial H\left(\boldsymbol{\theta}_{2\mathrm{DHDA}}\right)}{\partial \boldsymbol{\theta}_{2\mathrm{DHDA}}} = 2N\left(\boldsymbol{\theta}_{2\mathrm{DHDA}}^{\mathrm{T}} \boldsymbol{S}_b \boldsymbol{\theta}_{2\mathrm{DHDA}}\right)^{-1} \boldsymbol{\theta}_{2\mathrm{DHDA}} \boldsymbol{S}_b - \sum_{i=1}^{C} 2n\left(\boldsymbol{\theta}_{2\mathrm{DHDA}}^{\mathrm{T}} \boldsymbol{W}_i \boldsymbol{\theta}_{2\mathrm{DHDA}}\right)^{-1} \boldsymbol{\theta}_{2\mathrm{DHDA}} \boldsymbol{W}_i \tag{6}$$

However, there is no close-form solution for $\frac{\partial H\left(\boldsymbol{\theta}_{2\mathrm{DHDA}}\right)}{\partial \boldsymbol{\theta}_{2\mathrm{DHDA}}} = 0$. Instead, the gradient descent algorithm is used for the optimization of H and $\boldsymbol{\theta}_{2\mathrm{DHDA}}$ is solved. Usually, face images are projected onto the whole $\boldsymbol{\theta}_{2\mathrm{DHDA}}$, the most discriminant features are could not extracted, thus, former d column vectors of $\boldsymbol{\theta}_{2\mathrm{DHDA}}$ are selected as projection axes, then, the extracted features expressed as

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\theta}_{\text{2DHDA}}(:,1:d) \tag{7}$$

where $\boldsymbol{\theta}_{\text{2DHDA}}(:,1:d)$ denotes the former d column vectors of $\boldsymbol{\theta}_{\text{2DHDA}}$ and $\boldsymbol{Y}$ represents the extracted feature matrix of sample $\boldsymbol{A}$.

2.3 Nearest Neighbor classifier

After a transformation of 2DHDA, Nearest Neighbor classifier is selected to perform gender classification. Suppose $\boldsymbol{Y}_{\text{test}}$ denotes the feature matrix of an arbitrary testing sample $\boldsymbol{A}_{\text{test}}$, $\boldsymbol{Y}_j^i$ denotes the feature matrix of training sample $\boldsymbol{A}_j^i$. Then, the distance between $\boldsymbol{Y}_{\text{test}}$ and $\boldsymbol{Y}_j^i$ can be expressed as

$$D\left(\boldsymbol{Y}_j^i, \boldsymbol{Y}_{\text{test}}\right) = \sum_{k=1}^{d} \left\| \boldsymbol{Y}_j^i(:,k) - \boldsymbol{Y}_{\text{test}}(:,k) \right\|^2 \tag{8}$$

where $\boldsymbol{Y}_j^i(:,k)$ and $\boldsymbol{Y}_{\text{test}}(:,k)$ are the k th column vector of $\boldsymbol{Y}_j^i$ and $\boldsymbol{Y}_{\text{test}}$ respectively. $\left\| \boldsymbol{Y}_j^i(:,k) - \boldsymbol{Y}_{\text{test}}(:,k) \right\|^2$ is the Euclidean Distance between $\boldsymbol{Y}_j^i(:,k)$ and $\boldsymbol{Y}_{\text{test}}(:,k)$. If $D\left(\boldsymbol{Y}_q^p, \boldsymbol{Y}_{\text{test}}\right) = \min_{i,j} D\left(\boldsymbol{Y}_j^i, \boldsymbol{Y}_{\text{test}}\right)$ is satisfied, the testing sample $\boldsymbol{A}_{\text{test}}$ is classified to the p th class, where $\boldsymbol{Y}_q^p$ represents the feature matrix of training sample $\boldsymbol{A}_q^p$, and p , q are constants.

3. Experiments

3.1 Experimental objects

Experiments are based on Feret color face database and face database from University of Essex, UK. In Feret color face database, the images are varying in position, lighting and expression. We selected 10 male individuals, 10 female individuals with each individual 20 face images. Thus, there are 400 face images for experiments. In experiments, the images are chopped and resized to 100×90, then transformed to gray-scale images, as shown in Fig. 1.

In the face database from university of Essex, the images are with a resolution of 200×180, and with each individual 20 face images that vary in position, rotation, expression and lighting. We select 19 male and 19 female individuals, totally 760 face images for gender classification experiments. The original images are color images, we transformed them to gray-scale images and chopped them with a resolution of 80×70, as shown in Fig. 2.

3.2 Experimental results and analysis

In gender classification, there are only two classes, that are male and female respectively, thus in equation (4), $C = 2$. When the gradient descent algorithm is used for the optimization of H, $\boldsymbol{\theta}_{\text{2DLDA}}$ is selected as the initial matrix of $\boldsymbol{\theta}_{\text{2DHDA}}$ for iterations. Finally, Nearest Neighbor classifier is used for gender classification. The classification model is shown in Fig. 3.

Based on Feret color face database, firstly, the former 5 individuals of male and female, totally 200 face images are selected as training samples, the remains as testing samples. Experimental results are shown in Fig. 4. Secondly, for male and female, the former 4, 6, 8 individuals, totally 160, 240, 320 face images are selected as training samples respectively, experimental results are listed in Table 1.

Fig. 4 illustrates that, when totally 200 images are selected as training samples, the highest recognition rate of 2DHDA is 85.00%, which is 4.5% higher than that of 2DLDA. Table 1 shows that when 320 images are selected as training samples, the recognition rate of 2DHDA is 88.75%. However, the recognition rate of 2DLDA is only 83.75% and that of HDA is only 80.00%. When 160 and 240 images are selected as training samples respectively, we can know that the recognition rates of 2DHDA are also higher than that of 2DLDA and HDA. In table 1, when HDA is used for gender classification, PCA is used as a pretreatment step for dimensional reduction.

Based on face database from university of Essex, firstly, 20 individuals with 10 male and 10 female, totally 400 face images are selected as training samples and the remains as testing samples. When different numbers of feature dimension are selected, the results are shown in Fig. 5. Then, for male and female there are 9, 11 and 13 individuals for each class, totally 360, 440 and 520 face images are selected as training samples respectively, and the remains are selected as testing samples. Experimental results are listed in Table 2.

Fig. 5 demonstrates that, when 400 images are selected as training samples, the highest recognition rate of 2DHDA is 79.44%. The highest recognition rate of 2DLDA 70.89%, which is 8.55% lower than that of 2DHDA. Table 2 expresses that, the recognition rate of 2DHDA is higher than that of 2DLDA and HDA when 360, 440 and 520 images are selected as training samples respectively.

4. Conclusions and Future Work

In this paper, we presented the 2DHDA algorithm for gender classification using face images. Experimental results based on Feret color face database and face database from University of Essex show that 2DHDA is more effective than

2DLDA and HDA algorithms. However, when the images in a more complex background condition, how to improve the recognition rate need be further studied; the gradient descent algorithm is easy to run in local optimalization. Thus, in gender classification how to select the iteration factor and iteration number to gain the global optimal results need be further studied.

References

Gsles, M.J.F. (2002). Maximum Likelihood Multiple Subspace Projection for Hidden Markomodels[J]. *IEEE Transactions on Speech and Ausio Processing*, 2002, 10: 37-47.

Jing Wu, W.A.P., Smith, Eswin, R. Hancock. (2008). Facial Gender Classification Using Shape from Shading and Weighted Principal Geodesic Analysis[J]. *Springer link*, 2008, 5112: 925-934.

Marios, Kyperountas, Anastasios, Tefas, Ioannis Pitas. (2007). Weighted Piecewise LDA for Solving the Small Sample Size Problem in Face Verification[J]. *IEEE Transactions on Neural Networks*, 2007, 18: 506-519.

Modesto, Castrillon-Santana, Quoc, C. Vuong. (2007). An Analysis of Automatic Gender Classification[J]. *Springerlink*, 2007, 4756: 271-280.

Parinya, Sanguansat, Widhyakorn, Asdornwised, Somchai Jitapunkul, Sanparith Marukatat. (2006). Two-Dimensional Discriminant Analysis of Principal Component Vectors for Face Recognition[J]. *IEICE Transactions on Information and Systems*, 2006, 7: 2164:2170.

Saon, G., Padmanabhan, M., Gopinath, R., Chen. S. (2000). Maximum Likelihood discriminant Feature Spaces[J]. *IEEE International Conference on Acoustics, Speech, and Signal Processing,* 2000, 2: 1129-1132.

The face database, University of Essex, UK. http://cswww.essex.ac.uk/mv/allfaces/faces94.html

The Feret face database. http://face.nist.gov/colorferet/

Table 1. Experimental results based on Feret face database when different numbers of training samples are selected

Algorithms	160	240	320
2DHDA	84.58%	86.25%	88.75%
2DLDA	80.00%	85.62%	83.75%
HDA	76.66%	81.25%	80.00%

Table 2. Correct recognition rates based on face database from universityof Essex when different numbers of training samples are selected

Algorithms	360	440	520
2DHDA	74.50%	79.06%	77.50%
2DLDA	73.75%	70.31%	67.92%
HDA	65.75%	68.33%	69.16%

Figure 1. Face images in Feret color face database

Figure 2. Face images in face database from university of Essex

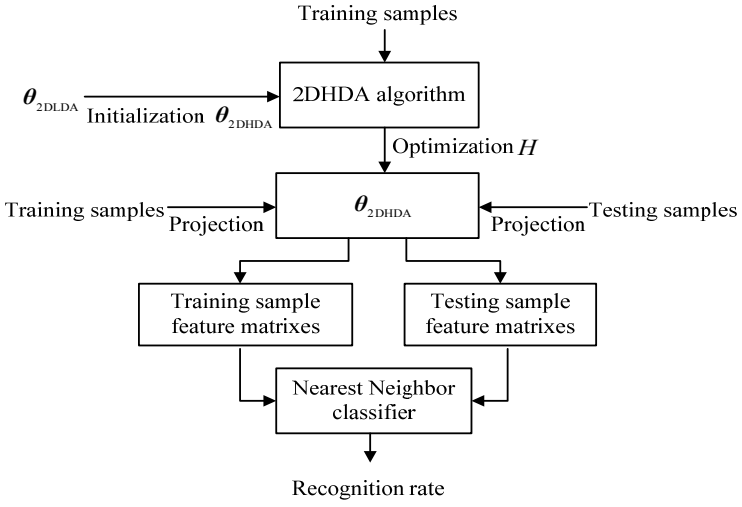

Figure 3. Gender classification model

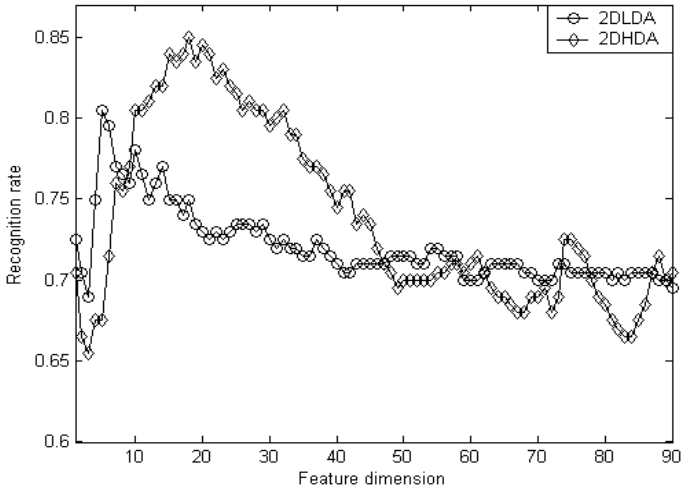

Figure 4. Experimental results based on Feret face database when 200 images selected as training samples

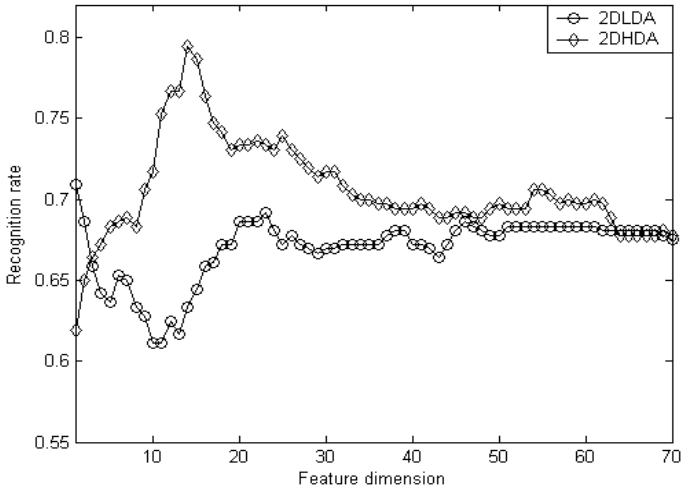

Figure 5. Experimental results with 400 training samples based on face database from university of Essex

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 2, No. 4

November 2009

Grid-based Data Quality and Data Integration Research

Xingman Li

College of Computer Science and Software, Tianjin Polytechnic University

Tianjin 300600, China

E-mail: simen_ok@163.com

Chunqing Li

College of Computer Science and Software, Tianjin Polytechnic University

Tianjin 300600, China

E-mail: frankly_lcq@163.com

Zhiyong Wang

Maton Information Technology Service (Tianjin) Co. Ltd

A303 Golden Sail Building, 18 Han Kou Xi Road, He Ping District, Tianjin 300057, China

E-mail: wangzy@tjmaton.com

Abstract

Data information is important strategic enterprise resources, rational and effective use of the correct data to guide business leaders to make the right decision-making, enhance the competitiveness of enterprises. The data quality and the data integration, very important speaking of the enterprise, is the enterprise innovation development power. In order to improve the efficiency of data integration, we must permit multiple applications to share computing resources. The grid technology may step isomerism platform computing resource to carry on the work distribution and to carry out, uses the existing hardware property effectively or new highly effective, the economical hardware, may carry on highly effective to the data, expands economically, so that adjustment and optimization face enterprise's data transmission.

Keywords: Data quality, Data integration, Data clean, Data grid, Data transmission, Isomerism platform

1. Background

In the present era, the enterprise informationization's request is getting more and more urgent, a very important aspect is the business data management. For most enterprises, ensure that data quality is a formidable challenge. PricewaterhouseCoopers issued the whole world data management survey result indicated. 75% of the companies believe that data lacking can lead to serious problems; over 50% of the companies overrun the cost due to the inner; over 33% of the companies can not but retard or give up use the new system's plan; over 20% of the companies thought that is unable to satisfy the contract or the agreement service level. [1] Up to 2009, Because of neglects the data quality question, some 50% above data warehouse project is unable to obtain the customer approval, even is defeated completely.

Although some projects might not overrun at all, good business planning requires taking this into consideration as part of the overall plan. It is a great challenge that must be faced to improve the data quality and reduce IT costs. The relation between improving data quality and data integration is interdependent. Improving data quality can make data integration more exact, whereas, we can improve data quality of a system with the help of data integration. Also, we can improve the data quality in the process of data integration. This both already may parallel, may also carry on separately (Ralph Kimball (2008)).

The main goal of gridding is to support coordinated work with the share source, which attribute to the result that the study on gridding data management has become very hot (Informatica (2008)). The effective and economic tensility for data integration can be achieved by developing the gridding computer system. Commercial hardware, for example, has shown the great demand for the tensility to reduce costs evidently. However, these griddings are dynamic for the nodes keep increasing and decreasing. Besides, collateral projects need to be timely adjusted in order to achieve the high-point, and reduce the load and frequency of modification of data mapping in order to answer the changing situation.

2. Data Quality Management

2.1 Summarization of Data Quality

In a very long time, data quality, also called intrinsic quality, mainly denotes the quality produced in the process of data production, such as precision, conformity, and integrality. The concept of data quality has been enlarged with the accumulation and widespread application. The contentment for the users' demands is the important target to scale the data quality.

In this meaning, data quality is a general term of a pair or a group of specific data precision, including the way of data input and flow in companies. Companies may not know the whole impact from the low or unknown data quality on the industrial operations if the definition of data quality is too narrow.

2.2 Standard for Measurement of Data Quality

2.2.1 Completeness

Measuring data needs elementary data elements, including correlative definitions and context-sensitive information for understanding and explaining data.

The process to create one master record may mean compiling or consolidating data from many records into single or multiple systems. Opportunities to perform consolidation or de-duplication must be considered (Keim DA, Panse C, Schneidewind J, Sips M, Hao MC, Dayal U. (2003)).

2.2.2 Consistency

There will always be challenges around alignment of data that need to be taken into account. Differences in data models across systems will lead to challenges with alignment of the structure of data and the actual data model to be used. Conformity can be used to measure whether tables in the database meet the specific regulations.

2.2.3 Conformity

Differences in data models across systems will lead to challenges with alignment of the structure of data and the actual data model to be used.

2.2.4 Integrity

It denotes the extent of information integrity, including entity integrity, citation integrity and domain integrity. Entity integrity requires each line in a table must be exclusive; citation integrity defines a cited relation between correlative rows in different tables of Relational Data System (RDS); domain integrity requests a row of data in the table needs to be in the legal data bound. The calculation of integrity is as follows: In the data set all satisfies the condition (to be possible to be above three one) in the data quantity/set records total *100%.

2.2.5 Mutuality

Most of the applications require accessing a certain data bound. To support specific applications, correctness is the essential attribute of data quality. Integrity, consistency and conformity all reflect the correctness in many aspects (Mike Schiff. Data Quality First(2006)).

Integrity tests the data correctness from the legal point of data numerical value; consistency relying on whether data accord with the application of logic; conformity from the lifecycle of the special product-data.

The relations between the characteristics of data quality are as follows: (Figure 1)

3. Data Integration Analysis

3.1 Summarization of Data Integration

Data integration is a process in which data from different sources and formats are integrated, logistical or physically. Data integration is traditionally divided into data warehouse and FDBS. Data warehouse technology centralizes data from many data sources into a central database physically. Whereas, only by interpreting users' queries into data sources queries, FDBS can integrate data logistically.

3.2 Federal Database System

FDBS is made of half-self-ruling database. Due to screening the differences from all kinds of the data sources, it can take a real-time and shortcut manipulation on the data of isomerous data sources (IQ Insight (2006)). Meanwhile, in FDBS, all the data sources provide the interfaces for the mutual accesses. FDBS can be centralized database, distributed database or other federal database.

3.3 Data Warehouse

An authority in the data warehouse filed called W.H.Inmon gave a short but comprehensive definition: data warehouse is a thematic, compositive and non-losable data aggregate, a decision-making process supporting management

department (Guo QJ, Yu HB and Wu K (2005)). In the process of enterprise management and decision-making, it is a thematic, compositive, time-relating and non-modificated data aggregate, in which data is classified as a generalized, function-independent and non-overlapped theme.

3.4 Middleware model

The above-mentioned methods, to some degree, solve problems the data-share and intercommunicative aspects, but at the same time, it also exists the following differences: FDBS mainly faces the integration among many databases, in which data sources may be mapped to every data mode. The enormous compositive system will endanger big difficulties in actual developing (Porto F, Silva VFV, Dutra ML, Schulze B.(2005)).

4. Informatica Data Quality and Data Integration Typical Solving Scenario Analysis

4.1 A Show of Scenario Flow Chart

Enterprises need to know more about the data in the source system. It is necessary for the enterprises to be able to integrate the data from many systems into a newer and more effective data intensity application procedure, as well as cleanse and enhance data.

To support today's business processes and goals, all corporate data needs to be universally accessible, flexible, reusable, and certifiably accurate (Mike Schiff. Data Quality First (2006)). Organizations need to know more about what is in their source systems, It is necessary for the enterprises to be able to integrate the data from multiple systems into new, more productive data-intensive applications, and they need to be able to cleanse and enhance data as well as monitor and manage the quality of data as it is used in different applications.

The platform of Informatica Data Integration and Data Quality provides the function of data analyzing, exploration, cleansing, conversion and coordinating, which can cooperate with each other in the different periods of data quality and data integration flow (Informatica (2008)). On this condition, it can provide the right and enterprise data quality service in the unification compositive environment. The concrete flow chart is as follow:(Figure 2)

4.2 FlowChartAnalysis

Data quality starts from the understanding of all the data in the system. According to the flow chart, we can see that accessing data, in batch and real-time modes, becomes especially important with the increasing data. Once the problems of data quality are found, it can be given a timely validation and correction to cleanse data, whereafter the well-cleansed data of high quality is transformed and reconciled to be integrated into the data system (Keim DA, Panse C, Schneidewind J, Sips M, Hao MC, Dayal U. (2003)). By doing that, it can make sure the data consistency as well as generating data monitoring report.

5. Conclusion and Outlook

There is no systemic appraisal target of data quality at present. The present data quality evaluation only aims at the important qualitative index, such as the problems of consistency, integrity and complexity. Data quality has become the increasingly serious problems which both big and small corporations are facing. It is vital to all the proposal of data integrity. The well data quality can furnish the corporations with supports on the decisions, becoming a new impetus for the corporations' innovation. Reversely, the cheesy data quality can bring unexpected resistance and even disaster. Improving data quality is a very long period pursuit; meanwhile, it needs incessant efforts and experiences. We should believe that data quality and data integrity will gain further development with maturity the Internet and cloudy calculation technology.

References

Guo, QJ, Yu, HB and Wu, K. (2005). "Research & application of distributed condition-based maintenance open system". Computer Integrated Manufacturing Systems, P416-421 (in Chinese with English abstract).

Hailong, Ting, and Hongbing, Xu. (2007). *Data quality analysis and application;* Computer technology and development NO.3 VOL.17;P1-3 Workshop, P 45-57.

http://www.businessobjects.com/pdf/products/eim /iq_insight.pdf.

Informatica (2008). "Making Data Work: Addressing Data Quality at the Enterprise Level"; Informatica White Paper; P2-4.

IQ Insight. (2006). "Data Quality Assessment Solution [EB/OL]".

Keim DA, Panse C, Schneidewind J, Sips M, Hao MC and Dayal U. (2003). *Pushing the limit in visual data exploration*: Techniques and applications. In: Günter A, et al., eds. P 37-51.

Mike Schiff. Data Quality First(2006)."It's Just Logical[R]".Business Objects White Paper.

Porto F, Silva VFV, Dutra ML and Schulze B.(2005). *An adaptive distributed query processing grid service.* In:

Pierson JM, ed. Data Management in Grids: 1st VLDB

Ralph Kimball. (2008). “An Architecture for data quality”. A Kimball Group White Paper.

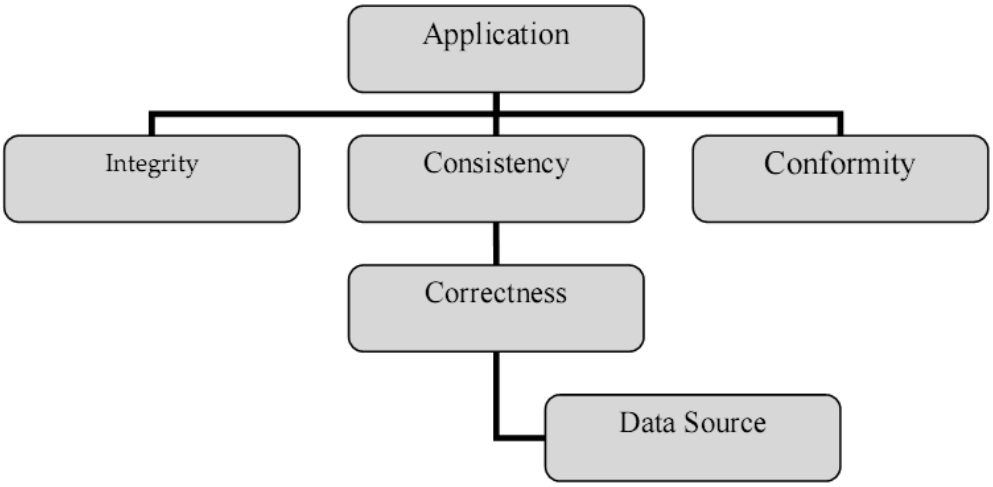

Figure 1. the characteristics of data quality chart

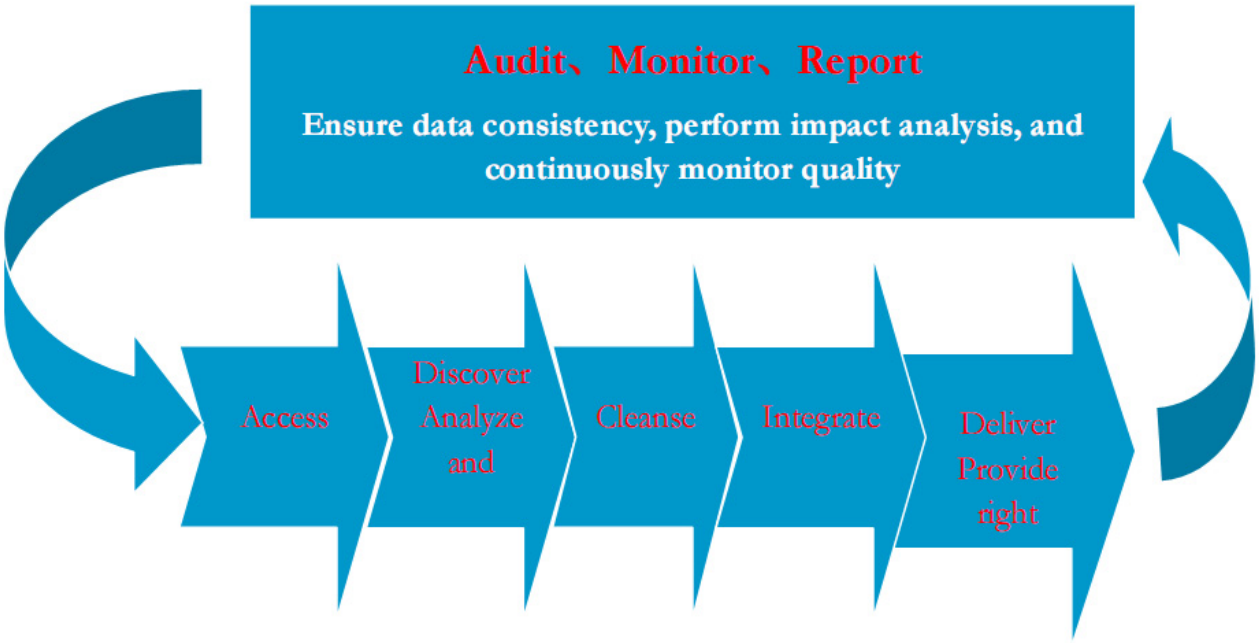

Figure 2. Scenario Flow Chart

A journal archived in Library and Archives Canada
A journal indexed in CANADIANA (The National Bibliography)
A journal indexed in AMICUS
A journal included in DOAJ (Directory of Open-Access Journal)
A journal included in Google Scholar
A journal included in LOCKSS
A journal included in PKP Open Archives Harvester
A journal listed in Journalseek
A journal listed in Ulrich's
A peer-reviewed journal in computer and information science

Computer and Information Science

Quarterly

Publisher Canadian Center of Science and Education
Address 4915 Bathurst St. Unit # 209-309, Toronto, ON. M2R 1X9
Telephone 1-416-208-4027
Fax 1-416-208-4028
E-mail CIS@ccsenet.org
Website www.ccsenet.org
Printer William Printing Inc.
Price CAD.$ 20.00

ISSN1913-8989