A Safe and Versatile Storage Server for Off-Site Backup
Hirofumi Amano, Yuki Dohi and Hiromune Ikeda

Evaluation of the Industrial and Social Impacts of Academic Research Using Patents and News Articles
Shumpei Linuma, Satoshi Fukuda, Hidetsugu Nanba and Toshiyuki Takezawa

A Proposal of Human and Machine Reminding for High Response Rate in Use of Web Questionnaire Survey System
Tetsuya Oishi, Masao Mori and Eiichi Takata

Expression Strength for the Emotional Scene Detection from Lifelog Videos
Atsushi Morikuni, Hiroki Nomiya and Teruhisa Hochin

An Effective and Appealing Instruction Based on Active Learning in a Large Class of Project-based Research Learning on Environmental Problems
Susumu Yamazaki, Tsukasa Morimoto, Tohru Futawatari, Atsushi Nogami, Taku Jiromaru, Masafumi Iwano & Tokuro Matsuo

Presumption Model for Postoperative Hospital Days from Operation Records
Takanori Yamashita, Yasuharu Nakashima, Naoki Nakashima, Yoshifumi Wakata, Yukihide Iwamoto, Sachio Hirokawa, Satoshi Hamai & Brendan Flanagan

Energy Efficient and Enhanced-type Data-centric Network Architecture
Shanming Zhang, Hidetoshi Takeshita, Satoru Okamoto & Naoaki Yamanaka
A Safe and Versatile Storage Server for Off-Site Backup

Hirofumi Amano¹
Kyushu University, Japan

Yuki Dohi²
e-Seikatsu Co., Ltd., Japan

Hiromune Ikeda³
CyberAgent, Inc., Japan

Abstract

This paper makes a proposal for a safe and versatile storage server for off-site backup based on secret sharing scheme and storage virtualization. This approach can save important data even after a disaster which destroys the primary data and some of its backup copies. It also prevents a secret from leaking to others. The design of the prototype utilizing the (3, 5)-threshold secret sharing scheme and iSCSI is described, and preliminary experiment is reported. Future enhancement plans are also discussed.

Keywords: Backup; secret sharing scheme; storage virtualization; iSCSI; disaster recovery

1. Introduction

Modern IT systems owned by an organization store its important and confidential data, such as citizenship records for a local government, account balances and transaction details for a bank, or medical records for a hospital. To avoid losing data by system troubles or operation mistakes, a backup copy of such vital data is replicated at regular intervals in separate media other than the one where the primary copy is stored. However, a disaster may destroy an entire IT system including the primary and backup copies of data stored in it. The Great East Japan Earthquake in March 2011 actually destroyed the citizenship records in several local authorities, and they had no choice but to substitute the lost data with the out-of-date snapshots maintained at the regional legal affairs bureau in a distant location [1].

The only solution to such a data loss is to maintain the up-to-date copy of data in a distant location. If we can keep the backup copy in our own computer installed in a remote place, we have no problem in restoring the important data even after a disaster. However, a small organization may not be able to find such a safe location under its control.

An internet data center (DC) may be a good choice for storing such backup data if we feel no fear in entrusting confidential information in the hands of outsiders. However, one problem of this approach is that the security policy of an organization may sometimes prohibit the transfer of important data out of the organization. Psychological factors like mere ignorance or groundless fear may magnify a negative opinion against DC solutions.

Therefore, encryption is often utilized to store vital data outside the control of its owner. Unfortunately, the encryption key must be used to restore the original data from the encrypted backup. The secret would be rendered vulnerable if the key must be stored in the same off-site backup facility. The security concerns and negative opinions against DC’s would become more serious, when we realize that we must duplicate encryption keys and store them outside our control to avoid data loss. Note that the encrypted data is technically equivalent to the secret itself if it is accompanied by the encryption key.

Meanwhile, our society has already introduced many IT applications running on various operating systems, which offer their own file systems having different access interfaces. If individual applications or OS’s require dedicated backup systems, the overall cost may become so high that small organizations cannot afford to introduce them. Therefore, the safe off-site backup system should be versatile, too. In other words, a single backup system should be capable of handling backup needs of various IT applications on various OS’s without modifying them.

To provide a safe and versatile storage system for off-site backup, we propose a mutual backup holding approach utilizing secret sharing scheme [2, 3] and storage virtualization technique [4]. In this approach, a group of organizations having a similar backup needs reciprocally look after backup copies of other organizations’ data in an encoded form.

The safety of such an off-site backup system consists of two properties: the robustness and the confidentiality. The robustness means that it must not lose the important data even after a large scale disaster. The confidentiality means that it must not leak the important information to others who have no proper authorization to know it. The versatility
means that the backup system can cope with various applications and OS’s without modifying them.

In this approach, the confidentiality and the robustness are achieved by the secret sharing scheme while storage virtualization technique gives us the versatility.

This paper reports the basic concept of the approach, the design of a prototype system, and its performance in preliminary experiments.

This paper is organized as follows. Section 2 describes two major technologies employed in our approach: secret sharing scheme and storage virtualization. Section 3 discusses some system design issues, including the requirement definition and several design decisions made for our prototype. It also gives an overview of the operation of the prototype. Section 4 reports preliminary experiment we performed with the prototype. Section 5 discusses related work. Future enhancement plans are discussed in Section 6. Section 7 gives conclusion.

2. Technological Elements

This section gives an overview of the basic technologies employed in the proposed approach, that is, secret sharing scheme and storage virtualization.

2.1 Secret Sharing Scheme

Secret sharing scheme (SSS) [2, 3] is a data encoding method which distributes a secret into a group of shares. The secret can be restored from a certain number of shares. Let k and n be the threshold and the total number of shares, respectively \( k \leq n \). A \((k, n)\)-threshold secret sharing scheme creates \( n \) shares from the secret. The secret can be reconstructed from at least \( k \) shares. However, it is absolutely impossible to restore it from shares fewer than \( k \). Fig. 1 shows an overview of \((3, 5)\)-threshold SSS.

Traditional encryption methods require an encryption key. To avoid the risk of data stealing, one should keep a single copy of the key. However, it leads to the loss of data when the only copy of the key is lost. If one must duplicate the key for a higher reliability, one must cope with a new problem: where and how one can store those copies safely. If a set of an encryption key and the encrypted data fall into a wrong hand, the secret will be leaked easily.

On the other hand, SSS requires no encryption key, and threshold \( k \) can be set higher when a higher security is necessary. To steal the secret, one must obtain \( k \) shares. A player who already has a share can restore the secret unjustly if and only if he/she can steal \( k - 1 \) shares from the other players.

Some SSS variations are based on simple Exclusive-OR (XOR) operations [5, 6], and can be processed faster than the other SSS variations. The algorithm proposed by Takaara et al. [6] introduces another parameter \( L \), which is the ratio of the size of the secret against that of the share. If \( L \) is set to 1, it is equivalent to the \((k, n)\)-threshold SSS based on XOR [5].

Let us introduce the following notations and definitions:

- \( S \) is the secret.
- \( \oplus \) denotes bit-wise XOR operation.
- \( A \parallel B \) denotes the concatenation of binary sequences \( A \) and \( B \).
- \( |S| \) denotes the bit length of \( S \).
- \( n \) is the number of the shares to generate. Let \( n \) be a prime number for simplicity\(^1\).
- \( k \) is the threshold where \( k \leq n \).
- \( d \) is a natural number such that \( |S| = (n - 1) \cdot d \). This implies a certain restriction on the choice of \( n \). For example, if \( |S| \) is set to 512, \( n \) can be 3, 5, 17, or 257.

The \((k, n)\)-threshold SSS based on XOR consists of two algorithms:

- the distribution algorithm which creates \( n \) shares from the secret data, and;

\(^1\) The distribution and recovery algorithms require \( n \) to be a prime number. However, we can extend the \((k, n)\)-threshold SSS for a composite number \( n_c \) without loss of generality by choosing a prime number \( n_p \) (\( n_p > n_c \geq k \)) and discarding \((n_p - n_c)\) shares after encoding.
the recovery algorithm which reconstruct the secret out of \( k \) shares.

The authors of [5] defined the distribution algorithm of the XOR-based \((k, n)\)-threshold SSS as shown in Table 1.

Table 1: Distribution Algorithm of the XOR-based \((k, n)\)-threshold SSS

| Input: | secret \( S \in \{0, 1\}^{(n-1)d} \) |
| Output: | shares \((W_0, \ldots, W_{n-1})\) |
| \( S_0 \in \{0\}^d \) | |

//[Step 1]
Divide \( S \) into \((S_1, \ldots, S_{n-1})\) equally;

//[Step 2]
For \((i = 0; i < k - 1; i++)\) {
  \( W_i = S_i \)
  For \((j = 0; j < n; j++)\) {
   Choose \( r^j \) from \( \{0, 1\}^d \) uniformly at random;
  }
Discard \( r^0_{n-1} \);

//[Step 3]
For \((i = 0; i < n; i++)\) {
  For \((j = 0; j < n - 1; j++)\) {
   \( W_{i,j} = S_j \)
   \( W_i = W_{i,0} \oplus \ldots \oplus W_{i,n-2} \)
  }
Return \((W_0, \ldots, W_{n-1})\)

An intuitive explanation of the distribution algorithm given in [5] can be summarized as follows.

First, the secret \( S \) is divided into \( n - 1 \) \( d \)-bit pieces \( S_1, \ldots, S_{n-1} \) at [Step 1].

Next, at [Step 2], \((k-1)n-1\) \( d \)-bit random numbers \( r^0_0, \ldots, r^0_{n-2}, r^1_0, \ldots, r^1_{n-1}, \ldots, r^{k-2}_0, \ldots, r^{k-2}_{n-1} \) are chosen from \( \{0, 1\}^d \).

At [Step 3], the pieces of shares are created using the following formula:

\[
W_{(i,j)} = \left( \bigoplus_{h=0}^{k-2} r^h_{i-h+j} \right) \oplus S_{j-1},
\]

where \( 0 \leq i \leq n - 1, 0 \leq j \leq n - 2 \) and \( S_0 \in \{0, 1\}^d \).

Finally, these pieces are concatenated to construct each share.

The authors of [5] outlined the recovery algorithm of the XOR-based \((k, n)\)-threshold SSS as shown in Table 2.

Table 2: Recovery Algorithm of the XOR-based \((k, n)\)-threshold SSS

| Input: | shares \((W_{t_0}, W_{t_1}, \ldots, W_{t_{k-1}})\) |
| Output: | secret \( S \) |

//[Step 1]
For \((i = 0; i < k; i++)\) {
  Divide \( W_i \) into \((W_{(i,0)}, \ldots, W_{(i,n-1)})\) equally;
}
Generate vector \( W_{(t_0, \ldots, t_{k-1})} \);

//[Step 2]
For \((i = 0; i < k; i++)\) {
  for \((j = 0; j < n - 1; j++)\) {
   Obtain binary vector \( v_{(t_i,j)} \) from indexes of \( W_{(i,j)} \);
  }
}

//[Step 3]
Generate matrix \( M^{(k,n)}_{(t_0, \ldots, t_{k-1})} \) by \( v_{(t_0,0)}, \ldots, v_{(t_{k-1},n-1)} \);

//[Step 4]
Execute Gauss-Jordan elimination on \( W_{(t_0, \ldots, t_{k-1})} = M^{(k,n)}_{(t_0, \ldots, t_{k-1})} \cdot R_{(k,n)} \);

//[Step 5]
Recover all divided pieces of the secret by the share combinations obtained at [Step 4].

Return \( S = S_1 \oplus \ldots \oplus S_{n-1} \);

An intuitive explanation of the recovery algorithm given in [5] can be summarized as follows.

Suppose that \( k \) shares \( W_{t_0}, W_{t_1}, \ldots, W_{t_{k-1}} \) are retrieved from the backup sites to reclaim the original data. The indexes \((t_0, t_1, \ldots, t_{k-1})\) of these shares are the ones designated by the distribution algorithm at the encoding phase.

First, each share is divided into \( d \)-bit pieces at [Step 1].

Next, at [Step 2], \((kn - 2)\)-dimensional binary vector \( v_{(t_i,j)} \) is obtained from indexes of \( W_{(i,j)} \). \( v_{(t_i,j)} \) is defined as the generator of \( W_{(i,j)} \) denoted by

\[
W_{(i,j)} = v_{(t_i,j)} \cdot R_{(k,n)},
\]

where \( R_{(k,n)} \) is a vector of unknown variables.
\[ R_{(k,n)} = (S_1, \cdots, S_{n-1}, r_0^0, \cdots, r_{n-2}^1, \cdots, r_0^{n-1}, \cdots, r_0^{k-2}, \cdots, r_0^{k-1})^T. \]

At [Step 3], the \((k(n-1) \times (k(n-2))\) binary matrix \(M_{(t_0, \cdots, t_{k-1})}^{(k,n)}\) is generated by \(v_{(t_0,0)}, \cdots, v_{(t_{k-1},n-1)}\) as follows:

\[ M_{(t_0, \cdots, t_{k-1})}^{(k,n)} = \begin{pmatrix} v_{(t_0,0)}, \cdots, v_{(t_{0},n-1)}, \cdots, v_{(t_{k-1},0)}, \cdots, v_{(t_{k-1},n-1)} \end{pmatrix}^T. \]

Therefore, all pieces of \(k\) shares can be combined in the following equation:

\[ W_{(t_0, \cdots, t_{k-1})} = M_{(t_0, \cdots, t_{k-1})}^{(k,n)} \cdot R_{(k,n)}, \quad (1) \]

where \(W_{(t_0, \cdots, t_{k-1})}\) is a vector of all pieces of shares denoted by:

\[ W_{(t_0, \cdots, t_{k-1})} = (W_{(t_0,0)}, \cdots, W_{(t_{0},n-2)}, \cdots, W_{(t_{k-1},0)}, \cdots, W_{(t_{k-1},n-2)})^T. \]

By executing Gauss-Jordan elimination with elementary row operation on \(GF(2)\) on the right-hand side of Equation (1), \(M_{(t_0, \cdots, t_{k-1})}^{(k,n)}\) is transformed into a matrix in the form of

\[ G_{(k,n)} = \begin{pmatrix} I & \Phi \\ \Phi & \Delta_k \end{pmatrix}, \]

where \(I\) denotes a \((n-1) \times (n-1)\) identity matrix, \(\Phi\)s are null matrices, and \(\Delta_k\) is a \((k-1)(n-1) \times (k(n-n-1))\) matrix.

Therefore, Equation (1) is transformed into the following form:

\[ W' = G_{(k,n)} \cdot R_{(k,n)}. \quad (2) \]

where \(W'\) is a vector obtained from \(W_{(t_0, \cdots, t_{k-1})}\) by the same Gauss-Jordan elimination operations.

By the definition of \(R_{(k,n)}\) and the form of \(G_{(k,n)}\), the first \(n-1\) elements of the right-hand side of Equation (2) must have the following structure:

\[ W' = \begin{pmatrix} S_1 \\ \vdots \\ S_{n-1} \\ \vdots \end{pmatrix}. \]

Since \(W_{(t_0, \cdots, t_{k-1})}\) is transformed into \(W'\) by the same Gauss-Jordan elimination, the first \(n-1\) elements of \(W'\) are obtained by XOR combinations of \(W_{(t_0,0)}, \cdots, W_{(t_{k-1},n-2)}\).

2.2 Storage Virtualization and iSCSI

The most important concept in storage virtualization [4] can be summarized as the unification of physically independent storage devices into a logical storage volume(s). A logical volume (LV) consists of a large number of data blocks, each of which can be accessed just by specifying its logical block address (LBA). An LV can span a number of disk drives, even a number of storage arrays or computers, each of which may contain multiple drives. The physical manifestation of a data block need not be the consecutive bits on a single drive. It can be striped over multiple devices, duplicated or recorded with redundancy for reliability reasons. A program accessing a logical data block in an LV need not know how the actual data bits are organized.

In a typical storage virtualization environment, a disk access is initiated by sending a SCSI command(s), and the response(s) from the device is sent back by a SCSI response(s) [7, 8]. Those commands and their responses can be transferred over various connections and networks.

For example, a high-end system may use a dedicated fibre channel (FC) storage area network [9] by encapsulating commands and responses in FC data frames, while a low-end system may employ its LAN by encapsulating commands and responses in IP packets. iSCSI [10,11] is one of the protocols to use the SCSI technology over the Internet. The maximum distance between an iSCSI storage server and its client is practically limitless.

In the SCSI/iSCSI terminology, a device/program sending a command and requesting the response is called an initiator, while the one receiving and responding to it is called a target. Fig. 2 illustrates the communication between an iSCSI initiator and an iSCSI target.

Fig. 3 illustrates the protocol stack for iSCSI from the viewpoint of a file system. Since the iSCSI protocol is implemented just below the SCSI layer, the file system software can access an external iSCSI target in the same way as its own internal SCSI disk drive. This means that an initiator can even format a target iSCSI logical volume just as it formats an internal SCSI disk drive.

The iSCSI initiator function is usually implemented as a software component. Most of recent OS’s have built-in initiators. On the other hand, the iSCSI target function can be implemented in various ways. A target can be a storage appliance equipped with a sophisticated hardware storage controller, or a simple computer running an iSCSI target software such as tgt [12, 13].
3. System Design Issues

This section discusses the design decisions we made for the prototype system.

3.1 Requirements on a Safe and Versatile Storage Server for Off-Site Backup

To minimize the additional investment for off-site backup facilities, we assume that a group of organizations having a similar off-site backup needs reciprocally look after backup copies of other organizations’ data. Moreover, we set the following requirements on an off-site backup system.

- **Robustness**: The data must be recoverable even after a disaster destroys an entire site together with some of its backup sites. This may sound too strict, but we should remember that some of the backup sites may also be devastated in case of a large scale disaster.

- **Confidentiality**: The contents of data must not be known to any persons other than its owner, including the system administrator of a backup site. The metadata (e.g. file name, size, ownership and timestamp) must not be known to others, either. The administrators of the backup sites must be able to perform their daily operations without knowing neither the contents nor the metadata of the deposited backup data.

- **Versatility**: The backup functions must be available to various applications on various OS’s and various file systems without modifying them.

3.2 The Prototype

To meet these requirements, the following design decisions have been made.

- The XOR-based $(k, n)$-threshold SSS is employed for encoding. In our prototype, $k$ and $n$ are set to 3 and 5, respectively. That is, the secret data can be reconstructed even after the original data and two of its shares are lost. This satisfies the robustness requirement.

- The backup data of an entire LV is saved in the form of $n$ shares of the same size. Every 512 byte block on the local LV is stored at the same LBA on the share LV’s after SSS encoding. These share LV’s are stored in $n$ remote sites. Any block update performed against the local LV can be reflected on its remote share LV’s by the corresponding block write operations of the SSS-encoded blocks at the same LBA. Since the entire volume is encrypted, even the remote site administrators can know neither the data nor the metadata in the volume, intentionally or unintentionally. The confidentiality requirement is hereby satisfied.

- We extend the functions of tgt [12, 13], an open-source iSCSI target on Linux, so that it automatically translates a single block-write operation into a sequence of a local block-write, SSS share generations for that block, and remote write operations of those share blocks via the iSCSI protocol. Since the iSCSI initiator features are already built-in most of the recent OS’s such as Unix/Linux/Windows, existing applications on these OS’s should have no problem in using our backup functions just by performing an ordinary iSCSI write operation against the extended iSCSI target (tgt-x). The versatility requirement is also fulfilled.

We chose the XOR-based $(k, n)$-threshold SSS instead of Blakley’s or Shamir’s algorithms, since it is most efficient. Table 3 shows the comparison of these three algorithms in terms of computational complexity where $S$ and $W$ are the secret and the share, respectively.
As shown in Table 3, the XOR-based \((k, n)\)-threshold SSS works more efficiently than Blakley's or Shamir's algorithms. Moreover, it can generate the share of the same size as the secret. This property is very convenient for our purpose, because we must maintain the same LBA space on the local LV and the share LV's. This is another reason why we chose the XOR-based \((k, n)\)-threshold SSS.

The performance advantage of XOR-based SSS can be confirmed by experiments, too. Fig. 4 and Fig. 5 show the comparison of the elapsed times for encoding 10 MB data by Shamir's algorithm and the XOR-based \((k, n)\)-threshold SSS algorithm for various \(k\) and \(n\). The XOR-based SSS clearly performs much better than Shamir's in all cases.

Fig. 6 shows an overview of our prototype. Needless to say, OS's exist also on all remote sites, but they are not shown in the figure for simplicity.

A client site must have a Linux storage server running a tgt-x. Applications may run on a different OS, such as Windows, if it is equipped with an iSCSI initiator.

A single block write call from the client is processed as follows, where Steps (1) through (4) are also indicated in the figure.

(1) An application on the client issues a block-write call against its local volume mounted via iSCSI. The built-in iSCSI initiator on the client OS delivers it to the tgt-x encapsulating it in an iSCSI PDU.

(2) The tgt-x performs the block-write operation on its local LV.

(3) It then generates \(n\) shares for that block and issues the corresponding block-write iSCSI PDUs against \(n\) targets on \(n\) remote sites.

(4) Each remote target performs the block-write operation requested by the tgt-x. Those share blocks are written at the same LBA as that on the local LV.

4. Preliminary Evaluation

This section describes the preliminary evaluation we carried out on the prototype, and reports the results.
4.1 System Configuration

To verify the functions and the performance, we prepared the environment shown in Fig. 7, which follows the same step-number notation employed in Fig. 6.

The computers were located on the same 1000Base-T LAN, but the network emulator was inserted to simulate a WAN environment where we can expect that some of the remote sites can survive a large scale disaster. The delay time, 22msec, matches the one we once observed between two sites located over 800km apart from each other.

The original plan for our off-site backup system is that a group of organizations having the similar off-site backup needs store the backup data mutually. It means that the n+1 participant sites must install the same tgt-x in this complete configuration. In the preliminary experiment, however, our main concern was to test its functionality and basic performance. Therefore, a tgt-x was installed in the local storage server only, and the off-site backup servers had a standard tgt without modification.

Before starting the tgt-x and the tgt’s services, a blank LV of the same size was created at each host on their local storage device and attached to the tgt-x and the tgt’s. After the tgt-x and tgt daemons started, an ordinary Linux file system was generated on the client’s local volume under the tgt-x, while the SSS-encoded versions were created on remote sites by standard tgt’s. Note that it is not necessary for the tgt-x to encode the entire volume before the file system creation, since any update made on the blank local LV will be reflected on the share LV’s, and the unused blocks in the local LV need not be recovered at the recovery phase.

The specifications of individual computers are given in Tables 4, 5, and 6.

4.2 Results

First, the functionality of our prototype must be verified. For this purpose, we examined the share LV’s on the “remote” backup servers. They actually contained some data, but they could not be mounted as a valid Linux file system. It is because they contained only the encoded version of a valid metadata. On the other hand, the contents of the local LV were reconstructed from three shares successfully. The functionality of the off-site backup system was hereby verified.

Second, we must evaluate its performance. To compare our approach with other methods, it is important to set fair conditions. To make the robustness constraint equal, the method to be compared must be able to survive a serious incident such that the original data and two of its backup...
data are lost. To implement this, we set up a simple 3-way NFS backup system which makes three full duplications for each file into three different NFS-mounted remote file systems on three of the emulated remote storage servers.

A test program performed simple write operations against the local storage server while changing the data size. Fig. 8 shows the write response times of the tgt-x and the 3-way NFS backup.

Although our prototype has handicaps of the encryption time and the greater number of remote hosts to access, it matches the 3-way NFS duplication method. It outperforms the 3-way NFS for data size over 2MB.

![Figure 8. Write Response Time of tgt-x and 3-way NFS Backup](image)

Read performance was not measured, since read operation requires neither remote access nor decoding. It is actually a simple read operation from the local storage server.

5. Related Work

This section gives an overview of the research work related to off-site backup.

5.1 Storage Replication Approach

DRBD (distributed replicated block device) [14] is a storage replication system implemented for Linux. It can replicate a local LV on another remote host also running DRBD. It is now widely used for backup purposes.

To meet the robustness requirement, we must run multiple DRBD’s in a daisy chain configuration. It will make the start-up sequence more complicated, and will not be suitable for situations where independent organizations mutually maintain backup data.

The confidentiality requirement cannot be satisfied by DRBD itself, since it has no encryption mechanism. If we use any conventional encryption scheme, however, we fall into the key store problem again.

DRBD operates at the block level, and we can create various file systems implemented on Linux. However, the versatility requirement cannot be satisfied by DRBD alone. To offer a backup service for systems other than Linux, it must have the assistance of another storage virtualization technique such as iSCSI.

5.2 File Sharing Approach

Distributed file systems, such as NFS [15] or CIFS [16], designed for file sharing can be used also for backup purposes. They can replicate important data on the file level. If each replicated file is encrypted properly, the confidentiality requirement is fulfilled partially. However, this file sharing approach cannot hide some additional information such as file ownership, timestamp, or directory hierarchy.

Moreover, we must note that the file sharing approach also leaves the key store problem unsolved.

6. Future Enhancement

The prototype leaves several points for enhancement. This section discusses some of the work we are currently working on.

6.1 Non Blocking and Deferred Write Operation

The prototype system reported in this paper performs any write operation only in a blocking manner. That is, it will not finish its execution until all the SSS-encoded shares are written on the remote share LV’s. This design has the following shortcomings.

- Response time leaves much to be desired. If the tgt-x can operate in a non-blocking mode, the response will be much faster.
- Fault tolerance is not considered very much. Since iSCSI has only limited capabilities for error handling, the tgt-x cannot endure serious communication failure other than loss of a few packets.
- System maintenance is not considered, either. This off-site backup system is designed for independent sites which should have their own plan for maintenance schedules. However, blocking-mode operation cannot permit the system down on any of the remote sites even if the system maintenance is mandatory.

Therefore, our next task will be concentrated on the non-blocking mode of the tgt-x write operation for performance improvement. Moreover, it must be able to endure a long-term system down on the remote sites. This new constraint will force us to introduce some logging mechanism which
can hold a series of transactions during the system down time of certain duration. Fig. 9 shows the basic design of the enhanced tgt-x capable of deferred write operation.

The write operation will be performed as follows.

(1) An application on the client issues a block-write call against its local volume mounted via iSCSI.

(2) The tgt-x performs the block-write operation on its local LV.

(3) It then generates $n$ shares for that block and stores them in its write log file.

(4) Later, the tgt-x (or another deferred write daemon) reads the block write sequence from the log, and performs the corresponding remote write operations through iSCSI.

The only question is whether the time-related information can be stored properly even if immediate update is performed on the local LV while deferred updates are performed on the share LV’s. Timestamps are created and managed only in the client OS, and they are embedded in block update operations just as the data. Therefore, deferred update using the log should have no problem in recording timestamps in the file system as far as the order of write operations is preserved.

6.2 Quick and Safe Recovery

Another direction of enhancement is quick and safe recovery from a disaster. We can think of the following three problems.

- It is important to note that the recovery process of secret sharing scheme requires the ID’s of the shares for restoring the lost data. If the secret data is lost in a disaster, we must tell a decoding program which share was retrieved from which site. If this small piece of knowledge is lost in the disaster, we will have a difficulty in the recovery.

- The current design only assures that the necessary information is maintained for recovering the secret after a disaster. It has no automatic recovery feature. Moreover, it has no security mechanism to ensure that the secret will not fall into the wrong hand.

- In the current scheme, the static recovery is assumed. That is, the entire LV must be recovered before restarting the service. However, on-demand reconstruction is also possible.

The simplest solution to the first problem is to distribute the share allocation information among all tgt-x’s to make sure that somebody can tell which share survived the disaster in which site, even if some of them are destroyed.

To cope with the second problem, we must introduce some authentication and authorization mechanism to the automated recovery. Considering the magnitude of damage and confusion to be caused by a large scale disaster, this authentication procedure cannot rely on IT-based methods only, but may be forced to incorporate some social engineering methods.

The third problem can be solved by on-demand and background reconstruction. Every read access to the lost volume can be processed by block-by-block recovery while other blocks are to be reclaimed in background process. This will make a significant improvement in recovery time, although the performance will be sacrificed until the reconstruction is completed.

7. Conclusion

In this paper, we proposed a safe and versatile storage server for off-site backup based on secret sharing scheme and storage virtualization. The prototype showed that it can safely store important data without losing it after a serious incident such that the data and two of its backup data are lost. It also showed that it matches the 3-way NFS duplication approach, and outperforms it for data size over 2MB, although our prototype has handicaps of the encryption time and the greater number of remote hosts to access.

Our approach can be used in situations where a group of organizations having the similar off-site backup needs cannot use DCs due to financial, administrational or psychological reasons. Since the secret will never be disclosed even to the system administrator, those organizations can entrust their important data to others and can hold the SSS-encoded data of other organizations without security or moral concerns. The additional cost...
involved is that for extra disk capacity which is getting cheaper and cheaper these days.

Even if one cannot find trustworthy partners for backup sharing, a single tgt-x and n data centers can do the same job. Since each piece of data to be deposited at the data centers is an SSS-encoded share, one can entrust those shares without entrusting an encryption key at the same data center. In this case, one can place n virtual machines having the normal iSCSI target capability, which is similar to our preliminary experiment environment.

Acknowledgment

This research was supported by JSPS KAKENHI Grant Numbers 23500048, 24300025, and 26330087. The authors would like to thank Mr. Kenta Kawaguchi, Mr. Kosuke Harada, and Mr. Satoru Miura for their assistance.

References


Hirofumi Amano is an associate professor in the Institute for Information Technology at Kyushu University. He also works in the Graduate School of Information Science and Electrical Engineering at Kyushu University. Dr. Amano’s research interest includes parallel processing, distributed processing, grid computing, and cloud computing. Dr. Amano has contributed to the development of the user account management workflow for the HPCI (high-performance computing infrastructure), which is the computational platform realized by connecting K computer and other major super computers in Japan with high speed back bone networks in order to accommodate versatile computational needs of users.

Yuki Dohi received his Master of Engineering degree in computer science from the Graduate School of Information Science and Electrical Engineering at Kyushu University in 2011. Mr. Dohi works as a software engineer at e-Seikatsu Co., Ltd., which offers a wide variety of cloud-based systems and services for the real-estate business in Japan. He is an expert of functional
programming languages. His current assignment at the company is the development of desktop applications for Windows.

**Hiromune Ikeda** received his Master of Engineering degree in computer science from the Graduate School of Information Science and Electrical Engineering at Kyushu University in 2013. He works as a software engineer at CyberAgent, Inc., which offers a wide variety of blog and entertainment services in Japan. Mr. Ikeda's professional expertise spans various aspects of server-side programming, including Java and node.js. His current assignment at company includes Ameba now (miniblog) and QJ (SNS dedicated for college students).
Evaluation of the Industrial and Social Impacts of Academic Research Using Patents and News Articles

Shumpei Iinuma\textsuperscript{a} \quad Satoshi Fukuda\textsuperscript{a} \quad Hidetsugu Nanba\textsuperscript{a} \quad Toshiyuki Takezawa\textsuperscript{a}
Hiroshima City University, Japan \quad Hiroshima City University, Japan \quad Hiroshima City University, Japan \quad Hiroshima City University, Japan

Abstract

In scientometrics and citation analysis, several measures for evaluating the industrial relevance or the impact of academic research fields have been proposed. What seems to be lacking, however, is that these measures could not evaluate the recent industrial relevance or impact of each field, because most of them rely on citations of research papers in patents and vice versa. In this paper, we attempt to evaluate the industrial and social impact of research fields using document classification techniques. Our method classifies research papers and news articles using systems including the International Patent Classification (IPC) and the KAKEN classification index. Then it evaluates the industrial and social impact of each field by comparing the number of research papers with the number of patents or articles in the IPC categories and the projects funded in each KAKEN category. In addition to the classification, we extracted key phrases of technologies to capture trends.

Keywords: scientometrics, text classification, patent, research paper, news article

1. Introduction

In scientometrics and citation analysis, researchers attempt to evaluate the impact of academic research or its industrial relevance. Such evaluations are intended to discover salient work or maintain academic quality; they are also leveraged to select research topics suitable for funding. The main evaluation method is peer review: other researchers in the field assess justifiability and evaluate the work from the academic aspect. However, the social and economic effects of research activities have gained increased importance, and it is necessary to capture these effects, which are difficult to evaluate with traditional peer review or citation analysis. To address this problem, evaluation of academic research should consider other aspects as well as the academic.

In this study, we classified research papers and news articles using several classification systems and then we evaluated the impact of each research field by comparing the number of papers with the number of patents or articles for each category. First, we classified research papers using the International Patent Classification (IPC) and examined the number of publications for each IPC category for a comparison with the number of patent applications. We also applied our automatic classification method to news articles to find research fields that attract social interest. We adopted a patent classification system because patents can be regarded as achievements of academic research and technical development. In addition to this examination, we tried to capture the trends in technologies by extracting phrases of elemental technologies from patents, research papers, and news articles.

A classification system for research fields plays an important role in peer review by research funding agencies, because reviewers are assigned according to the classification system. The KAKEN classification index was designed to classify projects by the KAKEN research fund in Japan. In addition to the analysis using patents and news articles, we classified research papers into the KAKEN categories, and compared the number of publications in each with the number of projects funded by the KAKEN research fund.

The remainder of this paper is organized as follows. Section 2 describes related work. Section 3 explains our method for the classification of research papers and news articles. Section 4 reports on the experiment, and discusses the results. We present some conclusions in Section 5.

2. Related Work

The purpose of this study is to evaluate the industrial and social impacts of academic research. This section describes related prior work. For the purposes of this study, we applied an automatic document classification method to documents that belong to different genres (research paper, patent, news article). This section also considers studies on cross-genre information access and automatic document classification.

2.1 Measurement of the Industrial Impact of Academic Research

Attempts to evaluate the industrial contribution of academic research have been based mainly on scientometric
analysis of relationships between industrial and academic research. A typical approach is to analyze citations between patents and research papers, as represented by Narin et al. [1]. They examined citations between research papers and patents in five countries: the United States, the United Kingdom, the former West Germany, Japan, and France. They regarded research papers and patents as science and technology respectively, and analyzed the impacts of science in each country on technology inside and outside the country. Such analyses using citations between patents and research papers have been done from various aspects, e.g., targeting specific fields such as laser medicine [2], space engineering [3], or focusing on publications and patents in specific locations [4].

However, a problem is that analyzing citations cannot always enable evaluation of recent impact. Cutting-edge work is not yet cited because propagation of knowledge takes time. In this study, we adopt an automatic document classification technique to solve this problem.

2.2 Cross-Genre Information Access and Automatic Document Classification

A great deal of effort has been devoted to cross-genre information access and document classification. An example is the task of technical trend research at NTCIR-3, which aims to retrieve patents related to a news article given as an input [5]. One of the points of this task was to take account of the difference of terms used in research papers and patents. For example, “resident” appears frequently in news articles, but not in patents. Therefore, if terms are weighted using inverse document frequency, the importance of a given term differs greatly between patents and news articles. Terminology differences between genres are also observable in research papers and patents. There are terms that appear only in patents, such as “document-editing device” which becomes “word processor” in research papers. Nanba et al. proposed a method to translate scholarly terms into patent terms using citations between patents and research papers or leveraging thesaurus automatically built from patents [6]. This method is useful when users search both research papers and patents in a particular field.

Another example is the Patent Mining Task of the Seventh and Eighth NTCIR Workshops [7, 8]. The task's goal was the creation of technical trend maps from a set of research papers and patents in a particular research field. The task was composed of the following two subtasks.

- Research paper classification: Classification of research papers in terms of the IPC. The aim of this subtask was to enable comprehensive collection of patents and research papers in the particular field.

- Technical trend map creation: Extracting expressions of elemental technologies and their effects from research papers and patents.

Most participant groups employed the k-nearest neighbor (k-NN) method in the subtask of research paper classification. This was because the IPC was designed minutely so that it has 30,855 categories at the lowest level. The number of training documents was also large, about four million. In this study, we examine the trends in research papers and news articles using the classification technique employed in the first subtask above.

Fukuda et al. proposed a method for research paper classification in terms of the KAKEN classification index [9]. They improved classification accuracy by focusing on the expressions of elemental technologies and their effects, and they also employed the k-NN method. We employed their method to classify research papers, then compare the number of documents with the selection of projects funded by the KAKEN research fund.

3. Classification of Research Papers and News Articles

The points we would like to make clear are as follows.

- Differences exist between the numbers of patent applications, research papers, and news articles in categories defined through the IPC system.

- Gaps exist between the numbers of research papers and of projects funded by the KAKEN research fund in many research fields.

For the purpose of the first analysis, we classified research papers and news articles into the IPC system using a classifier based on the k-NN method. The IPC system is a global standard hierarchical patent classification system, which is organized as a five-level hierarchy: section, class, subclass, main group, and subgroup. Figure 1 and Table 1 show examples for the IPC code “G06F 1/32”. In this study, we classified the documents into categories at the third level according to the 8th version of the IPC system (subclass level; 643 categories).

```
<table>
<thead>
<tr>
<th>G</th>
<th>06</th>
<th>F</th>
<th>1</th>
<th>/32</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subclass</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Main group</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Figure 1. Hierarchy of the IPC system using the example “G06F 1/32”
Table 1. IPC code example for “G06F 1/32”

<table>
<thead>
<tr>
<th>IPC code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>Physics</td>
</tr>
<tr>
<td>G06</td>
<td>Computing; calculating; counting</td>
</tr>
<tr>
<td>G06F</td>
<td>Electrical digital data processing</td>
</tr>
<tr>
<td>G06F 1</td>
<td>Details of data-processing equipment not covered by groups G06F 3/00 to G06F 13/00</td>
</tr>
<tr>
<td>G06F 1/32</td>
<td>Means for saving power</td>
</tr>
</tbody>
</table>

Table 2. Examples from the KAKEN classification index

<table>
<thead>
<tr>
<th>Area</th>
<th>Discipline</th>
<th>Research field</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informatics</td>
<td>Computing</td>
<td>Computer system, software, information network, etc.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Human informatics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cognitive science, soft computing, human interface and interaction, etc.</td>
</tr>
<tr>
<td>Social science</td>
<td>Economics</td>
<td>Theoretical economics, finance, economic history, etc.</td>
</tr>
<tr>
<td></td>
<td>Psychology</td>
<td>Social psychology, educational psychology, clinical psychology, etc.</td>
</tr>
</tbody>
</table>

The classifier employs a patent retrieval system that indexes nouns, verbs, and adjectives, and adopts Okapi BM25 as its similarity measure [10]. First, the top $k$ documents $\{d_1, d_2, \ldots, d_k\}$ with highest similarity (k-NN) are retrieved for a research paper or a news article given as input. When the system computes the similarity between a research paper and a patent, it uses the title of the research paper and a patent, respectively. The full text of research papers was not used. The similarity is computed as $\text{sim}(\text{title}(p), \text{spec}(d)) + \text{sim}(\text{author}(p), \text{inventor}(d)) + \text{sim}(\text{venue}(p), \text{spec}(d))$, where $p, d$ refer to the research paper and the patent respectively. The full text of research papers was not available with our dataset. For news articles, the system computes the similarity between a news article and a patent as $\text{sim}(\text{body}(a), \text{spec}(d))$ where $a, d$ refer to the news article and the patent, respectively. Then, the system calculates $\text{score}(c)$ for the IPC codes of the retrieved documents. Here, $\text{score}(c)$ can be regarded as a measure of the likelihood that the input document has label $c$ (IPC code). Finally, the input document is classified into the IPC code with the highest score. Our system uses the following ranking method (the Listweak method) [11]:

$$\text{score}_{\text{Listweak}}(c) = \sum_{i=1}^{k} \text{occur}(c, d_i) \text{sim}(q, d_i) r^i$$  \hspace{1cm} (1)

where $\text{occur}(c, d_i)$ returns 1 if document $d_i$ has IPC code $c$, otherwise it returns 0. $\text{sim}(q, d_i)$ denotes BM25 similarity between input $q$ and $d_i$ (a document retrieved by the patent retrieval system). In addition, $r^i$ is a penalty factor against documents with lower rank, $r$ is set to 0.95 in our system. $r$ was determined by preliminary experiments.

For the classification into the KAKEN classification index, we used the method proposed by Fukuda et al., which focuses on the expressions of elemental technologies and their effects [9]. Their method is also based on the k-NN method. The KAKEN classification index is a classification system designed by the KAKEN research fund, and it is used to determine resource distribution. The index is organized as a three-level hierarchy: Area, Discipline, and Research Field, and it has been modified as new research fields arise. Table 2 shows examples from the index.

The classification method proposed by Fukuda et al. is characterized by its indexing module. They set weights on words, using a key phrase list extracted in advance. Some studies have shown that adjusting the weights on words appearing in documents is effective for classification [12, 13]. Fukuda et al. extracted expressions of elemental technologies and their effects from research papers as key phrases. They used the information extraction method developed in their prior study [14]. The extraction method is based on machine learning and formulated as a sequence-labeling problem. An example of key phrase extraction is given in Figure 2. The information extracted is defined as follows:

- **TECHNOLOGY**: Expressions about algorithms, materials, tools, and data used in studies;
- **EFFECT**: Pairs of ATTRIBUTE and VALUE;

Through &lt;TECHNOLOGY&gt;closed-loop feedback control&lt;/TECHNOLOGY&gt;, the system could &lt(EFFECT&gt;&lt;VALUE&gt;minimize&lt;/VALUE&gt;&lt;ATTRIBUTE&gt;&lt;ATTRIBUTE&gt;&lt;/EFFECT&gt;)

Figure 2. Example of key phrase extraction. (Translated from Japanese)

They employed the support vector machine approach, which obtained higher precision than the conditional random field [15] approach. They conducted an experiment using the dataset for the NTCIR-8 Patent Mining Task. They obtained recall and precision scores of 0.276 and 0.539, respectively. Key phrases extracted were used to set the weights on word frequencies appearing in documents, and they studied the weights’ effect on the similarity when the system retrieves k-NN documents. Fukuda et al. also adopted the Listweak method, which we described earlier in this section. In addition to their classification method, we leveraged the key phrase extraction technique itself to examine which technologies were mentioned frequently in each document set: patents, research papers, or news articles. Gao et al. built a model to calculate the technology life cycle based on various patent-related indicators, such as the number of unique inventors and citations [16]. Phrases extracted in this study could also be useful as indicators of life cycles of technologies.
4. Experiments

First, we built the document classifiers, then classified research papers and news articles into the IPC system and counted the number of documents for each IPC code. We also classified research papers in terms of the KAKEN classification index, and compared the number of documents with the number of projects funded by the KAKEN research fund.

4.1 Building and Evaluating the Classifier

We built classifiers for research papers using the IPC system. Table 3 shows the patent data used for the k-NN method described in the previous section. All documents in the Japanese published unexamined patent applications and Unite States Patents have IPC codes that are manually assigned. Each set of documents is used in the classifier for research papers written in Japanese or English respectively. Next, we evaluated the classifiers using the test collections for the research paper classification subtask at NTCIR-7, -8. The task contains four subtasks:

- English subtask: Classification of English research papers using patent data written in English.
- Cross-lingual subtasks (J2E, E2J): Classification of Japanese research papers using patent data written in English, and vice versa.

We evaluated the classifier for Japanese research papers using the test collection for the Japanese subtask. Similarly, we evaluated the classifier for English research papers using the test collection for the English subtask. Each classifier was evaluated by the precision of the first ranked classification code. It should be noted that we classified research papers without using their abstracts, although the documents in the test collections contain title, authors, source, and abstract.

4.2 Classification into the IPC System and the KAKEN Classification Index

Table 4 shows the data that we classified for the analysis. We used JST\(^1\) Scientific and Technological Data during the period 2003-2012, articles from Yomiuri News that mention development or practical application of technologies (1993-2012), and news articles from TechCrunch\(^2\), which is a news website, focused on information technology (June 2005-December 2013). It should be noted that JST Data is bibliographic data, therefore, abstracts or the full text of research papers were not available. We used title, authors’ names, and publication name for classification. News articles from Yomiuri and TechCrunch contained the headline and body of each article. The number of documents in the Japanese published unexamined patent applications for each IPC category was used for a comparison (1993–2012). We classified JST Scientific and Technological Data and Yomiuri News articles into the IPC system using the classifier for research papers written in Japanese. In the same way, articles from TechCrunch were classified using the classifier for research papers written in English. In addition, we classified JST Data in terms of the KAKEN classification index, using the k-NN based classifier developed by Fukuda et al. Their classifier obtained a precision score of 0.827 on classification by the title of research papers.

<table>
<thead>
<tr>
<th>Data</th>
<th>Language</th>
<th>Classification system</th>
<th>Num. of docs</th>
</tr>
</thead>
<tbody>
<tr>
<td>JST Scientific and Technological Data</td>
<td>Japanese</td>
<td>IPC</td>
<td>850k</td>
</tr>
<tr>
<td>(bibliographic data)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yomiuri News (Articles about development or practical application of technologies were selected.)</td>
<td>Japanese</td>
<td>IPC</td>
<td>8,674</td>
</tr>
<tr>
<td>TechCrunch (IT news)</td>
<td>English</td>
<td>IPC</td>
<td>120,596</td>
</tr>
</tbody>
</table>

4.3 Extraction of Key Phrases

We extracted expressions of elemental technologies from the data described in previous subsections including Japanese Patents, JST Data, and Yomiuri News. We focused on documents belonging to a specific field that has a large population. The object field was chosen according to the results of the classification described in previous subsections. The purpose of the extraction was to find active technologies and capture technological trends through these documents. We employed the method developed by Fukuda et al. to extract and count expressions. In addition to this analysis, we

---

1. Japan Science and Technology Agency: http://www.jst.go.jp/
examined the numbers of expressions about specific technologies for each year. We chose technologies that frequently appear in both Japanese Patents and JST Data. Then we compared the trends in patents and research papers.

5. Results

As described in Section 4.1, the classifiers were evaluated using the test collections for the task at NTCIR-7, -8. For the results of evaluation we obtained precision scores of 0.815 (at $k = 300$) and 0.656 (at $k = 300$) on the Japanese and English paper classification, respectively. Examples of classified documents are shown in Figure 3 and 4. As can be seen from these figures, both documents were classified correctly.

Table 5 shows the number of documents in JST Scientific and Technological Data captured through the IPC system. The number of documents classified as G06F (Electrical digital data processing), A61K (Preparations for medical, dental, or toilet purposes), and C12N (Microorganisms or enzymes) are relatively large. It is clear that there are a large number of studies in these fields, as seen through the IPC system.

Table 5. The number of research papers for each IPC code (top 10).

<table>
<thead>
<tr>
<th>IPC</th>
<th>Description</th>
<th>Num. of docs</th>
</tr>
</thead>
<tbody>
<tr>
<td>G06F</td>
<td>Electrical digital data processing</td>
<td>94,943</td>
</tr>
<tr>
<td>A61K</td>
<td>Preparations for medical, dental, or toilet purposes</td>
<td>65,059</td>
</tr>
<tr>
<td>C12N</td>
<td>Microorganisms or enzymes</td>
<td>62,034</td>
</tr>
<tr>
<td>G01N</td>
<td>Investigating or analyzing materials</td>
<td>49,391</td>
</tr>
<tr>
<td>H01L</td>
<td>Semiconductor devices</td>
<td>42,070</td>
</tr>
<tr>
<td>H04N</td>
<td>Pictorial communication, e.g., television</td>
<td>18,847</td>
</tr>
<tr>
<td>G02F</td>
<td>Optical operation</td>
<td>17,428</td>
</tr>
<tr>
<td>A61B</td>
<td>Diagnosis; surgery; identification</td>
<td>15,845</td>
</tr>
<tr>
<td>C01B</td>
<td>Nonmetallic elements</td>
<td>15,613</td>
</tr>
<tr>
<td>A01G</td>
<td>Horticulture</td>
<td>12,728</td>
</tr>
</tbody>
</table>

Next, Table 6 shows the numbers of applications in Japanese Patents, research papers (JST Data), and news articles (Yomiuri News) for each IPC category. Some categories such as A23L (Food, foodstuffs) and A61K (Preparations for medical, dental, or toilet purposes) appear at a relatively higher rank in terms of the number of news articles. Specifically, A01G (Horticulture) appears in the top 10 list of news articles, but not of other genres. According to the IPC, Section A is defined as “Human necessities.” This result indicates that research and development on daily necessities tends to attract public concern. Comparing patents and research papers, there is some variability in IPC categories appearing in the results. Thus we see that the briskness of academic research, as captured by the number of publications, does not always lead to patent applications. IPC categories such as G06F (Electrical digital data processing), A61K (Preparations for medical, dental, or toilet purposes), and H04N (Pictorial communication, e.g., television) were ranked in the top 10 with regard to all genres. From these results we can say that research in these fields is likely to lead to patent applications, which we regard as an achievement of research and development, and tends to attract social concern. While we do not show the results in tables, we classified articles from TechCrunch using the IPC system. It should be understood that about 90% of these articles were classified into G06F (Electrical digital data processing) or G06Q (Data processing system or methods).
Figure 5 shows proportions of the number of documents in each KAKEN category. It shows the proportions for each data set, KAKEN-funded projects and JST Data. The results were obtained from the classification of research papers (JST Data) using the KAKEN classification index. Each document was classified at the third level of the index. The figure shows the result of classification at the highest level (Area level), because the number of categories at the third level is so large. While the proportion of “Engineering” is highest in the JST Data, “Medicine” makes up the largest proportion of KAKEN-funded projects. It is clear that research funds such as KAKEN put higher value on “Medicine,” even if the briskness of research on “Engineering” is the highest as the number of documents indicates.

Next, we describe the results of key phrase extraction. We counted the number of appearances of key phrases (TECHNOLOGY) extracted from documents classified as “G06F.” We used this category because the number of documents belonging to it is relatively large (see Tables 5 and 6). Table 7 shows the 10 most frequent phrases tagged as TECHNOLOGY in each document set. The sources of phrases are documents written in Japanese, but the table shows translations. They were extracted from sections describing “effect of invention” in patents, titles of research papers, and headline and body of news articles. As can be seen from the Japanese Patents column, more general terms appear in patents than in research papers (JST Data). The reason for this is not hard to see: the terms used in patents are often more abstract or creative than those used in research papers, because they are intended to widen the scope of claims. We will show the results of analysis focusing on some specific phrases later. Because nonexperts write most news articles and their readers are also nonexperts, phrases extracted from news articles are also more abstract. Some of them are familiar in daily life, e.g., e-mail, microwave, and mobile computer. With regard to research papers, the top two in the list are related to machine learning: neural networks and genetic algorithms. We can observe the high research activity in these areas.
In the same way, we count the number of appearances of phrases tagged as ATTRIBUTE, which we described in Section 3. We targeted news articles to study social trends: what kind of effects people expect from technologies. Table 8 shows the phrases that frequently appeared in news articles. The five most frequent phrases were cost, precision, safety, consumed electricity, and capability. It seems reasonable that examining effects of technologies enables us to see social expectations from technologies. However, object technologies are not mentioned in Table 8. We explain some of these object technologies. For example, take “cost,” which has been mentioned in news articles about various technologies’ effects on “cost,” e.g., a new method for Freon destruction or reuse of waste materials. With regard to “consumed electricity,” the number of articles related to televisions or computers was relatively large.

Table 8. Ten most frequent phrases of ATTRIBUTE in news articles classified as G06F. (Translated from Japanese)

<table>
<thead>
<tr>
<th>Phrase (ATTRIBUTE)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost</td>
<td>50</td>
</tr>
<tr>
<td>Precision</td>
<td>47</td>
</tr>
<tr>
<td>Consumed electricity</td>
<td>43</td>
</tr>
<tr>
<td>Capability</td>
<td>43</td>
</tr>
<tr>
<td>Durability</td>
<td>35</td>
</tr>
<tr>
<td>Fuel consumption</td>
<td>30</td>
</tr>
<tr>
<td>Reliability</td>
<td>25</td>
</tr>
<tr>
<td>Quality</td>
<td>25</td>
</tr>
<tr>
<td>Performance</td>
<td>20</td>
</tr>
<tr>
<td>Production cost</td>
<td>17</td>
</tr>
</tbody>
</table>

To capture the trends in patents and research papers, we chose three phrases: Neural Network, Carbon Nanotube, and Optical Fiber. We chose these because these phrases appeared frequently in the Japanese Patents and research papers datasets. While we used documents belong to “G06F” in the previous analysis, all of the documents were employed in this analysis. The number of phrases extracted from news articles is relatively small in comparison with other datasets, so we used only patents and research papers. We counted the number of appearances for each extracted phrase year by year. It should be noticed that we count only extracted phrases, which is different from just counting all appearances of the phrases in documents. First, Figure 6 shows the frequency for Neural Network. The number of uses of Neural Network as a technology decreases with respect to research papers. While the frequency of the phrase is high in research papers, the frequency in patents is low. We may say that Neural Network itself did not contribute to industry, but it has attracted the interest of researchers. Figures 7 and 8 show the frequency for Carbon Nanotube and Optical Fiber, respectively. There is a common pattern of increase and decrease in both patents and research papers, especially in Figure 8 (Optical Fiber). There are common peaks in 2006 and 2009. It is fair to say that the extent of academic research affects development of technology (patents) and vice versa.
5. Conclusion

In this study, we demonstrated that classification using various systems enables comparison between research fields from alternative perspectives. We classified research papers and news articles using the IPC system and then compared the numbers of documents with those of patent applications for each field. The results showed that the number of news articles mentioning developments and practical applications in fields related to food or medicine is relatively large, so these fields tend to attract social concern. In addition we classified research papers in terms of the KAKEN classification index which was designed by the KAKEN research fund, and compared the proportions of documents in each category with those of funded projects by KAKEN. From the results, we identified some gaps between the briskness of research and the number of projects funded by KAKEN research fund.

To examine social concerns, we used news articles. It should be realized that most of the articles about information technology (TechCrunch) were classified into categories related to data processing. For further analysis, classification at a finer level is required. As described in the previous section, we achieved a precision score of 0.656 on the classification of English research papers at the subclass level. Given the limitations of our classifiers, it is difficult to classify at the group level. The classifiers built in this study used only title, authors, and source of publications, because JST Data do not include the abstract or full text. If the abstracts or full text of research papers were used, classification performance should improve. In addition to the data-specific problem, the content difference and terminology difference between patents and news articles should be taken into account. In news articles, technologies tend to be described by focusing on superficial characteristics such as practical applications, although patents mainly describe the technologies themselves. Addressing these problems should lead to alignment of news articles with specific patents, and should enable further analysis.

In addition to the document classification, we employed an information-extraction technique to capture technology trends in industry, academia, and society. We studied patents, research papers, and news articles, respectively. As a result, we showed the top 10 frequently mentioned technologies used in each document set, and the activity in research fields. However, phrases extracted from patents were abstract and creative; therefore, we could not see specific technologies in patents as we could in research papers. To discover more-specific technologies, we should prepare some filters against generic terms. We also examined the frequency for each year, focusing on specific phrases. The results showed that activity in academic research affects development of technology, and vice versa. It should be also added that the extracted phrases including elemental technologies and their
effects could be used as indicators for life cycle analysis of technologies. For example, we could detect the reason for peaks by using phrases tagged as EFFECT, including VALUE andATTRIBUTE. These studies help us to see expectations in technologies. We plan like to study this aspect further in our future work.

Acknowledgments

This research owes much to the scientific and technological data of the Japan Science and Technology Agency. I would like to thank JST for providing the data.

References


Shumpei Iinuma is a graduate student in the Graduate School of Information
Satoshi Fukuda is a Ph.D. student at the Graduate School of Information Sciences, Hiroshima City University. He graduated from the Faculty of Information Sciences, Hiroshima City University in 2011, and the master’s program at the Graduate School of Information Sciences, Hiroshima City University in 2013. His research interests include natural language processing and information retrieval.

Hidetsugu Nanba is an associate professor in the Graduate School of Information Sciences at Hiroshima City University. Dr. Nanba’s research focus is natural language processing, text data mining, and information retrieval. He received his Ph.D. from Japan Advanced Institute of Science and Technology in 2001. He is a member of the Association for Computing Machinery (ACM) and the Association for Computational Linguistics (ACL), Information Processing Society of Japan (JSAI), the Association for Natural Language Processing, and Information Processing Society of Japan (IPSJ).

Toshiyuki Takezawa is a professor in the Graduate School of Information Sciences at Hiroshima City University. Dr. Takezawa’s research interests include natural language processing, speech translation and dialogue systems. He received his B.E., M.E., and D.Eng. degrees in electrical engineering from Waseda University in 1984, 1986, and 1989. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE), Information Processing Society of Japan (IPSJ), Japanese Society for Artificial Intelligence (JSAI), Acoustical Society of Japan (ASJ), and the Association for Natural Language Processing.
A Proposal of Human and Machine Reminding for High Response Rate in Use of a Web Questionnaire Survey System

Tetsuya Oishia, Masao Moriga, Eiichi Takataa
Kyushu University, Japan

Abstract

We held a web questionnaire survey to obtain some educational information of the students. During the term of holding the survey, we sometimes reminded the students to answer the questions. Moreover the targets who should be sent reminders were also promoted to answer by the teachers. We found that the reminders had a good effect. Actually, the response rate was 13.6% before the first reminder, but it became to be 36.6% after three reminders. In this paper, we describe this effect.

Keywords: Institutional Research, Questionnaire Survey.

1. Introduction

Institutional research is an important activity to plan ahead, decide something, and improve itself [1]. Japanese national universities have to report their activities to the government. These reports include some information such as educational effects, outcomes of research, and so on. Every university is evaluated by experts through the reports. If the evaluation is good, the university is regarded as a good institution in Japan or all over the world.

To show some advantages, the staffs who belong to university have to obtain the evidences demonstrating some effects or outcomes. In our university, the Researcher Database system is working to self-check and store some data for showing teachers' outcomes. But we don't have a system to obtain data from students our whole university.

So we introduced a web questionnaire system (we will mention in Section 4 to obtain students' information. Using this system, we could know how long the students studied in a day, whether they used the syllabus, or other activities. Basically the students had to answer some questions, but they weren't imposed any penalties even if they didn't answer any questions. Moreover, we didn't give any rewards to the students who answered completely.

Generally, there are few answers obtained from such a method with no rewards. In an experimental study about a survey response rate [2], the rate of those who weren't given any rewards was 17.8%. On the other hand, the rate of those who were given some rewards was 36.7%. In other words, the survey response rate depends on whether there are some rewards or not.

Our survey response rate was finally 36.6% without rewards or penalties. We reminded the students several times to answer a questionnaire. At the same time, the teachers of these students worked to remind them to answer. If we held the questionnaire by paper, it would be difficult to remind them to answer. But we used a web system, so it was very easy to do it. We confirmed that the completed answers increased after we reminded the students to answer. In this paper, we show the process to get some completed answers using reminders.

In our university, the questionnaire surveys had conducted to know the life of the students in 2007 [3] and 2011 [4]. The former was conducted by paper and the latter was conducted by web. Moreover the answerers of the former questionnaire were selected by random sampling. The response rates were 45.0% (2,299/5,107) and 19.6% (3,195/16,342) respectively. The response rate of web questionnaire was smaller than the one of paper questionnaire. But the number of answers was opposite. The web questionnaire can let many students answer, then the number of answers also increase. So we thought that the web questionnaire was very important tool. In addition, for the survey this time, we utilized the same web survey system, LimeSurvey [10], as the case of student survey in 2007 [3], using e-mail reminding function. It is interesting to see that 36.6% response rate in 2013 are significantly different from 19.6% response rate in 2007, where p-value is $2.2 \times 10^{-16}$.

We have already described this questionnaire survey simply [5] [6]. We showed only a fact obtained from conducting this survey in these papers. In this paper, we additionally show the backgrounds and the related works of our research. Then we show a reliability of the effect of reminders.

In the following parts, first we introduce our questionnaire in Section 2. Then we show the paper questionnaire in Section 3 and the Web questionnaire system in Section 4. In Section 5, we show how to remind the answerer comparing paper questionnaire and Web questionnaire. Section 6 shows the effects of reminders and we conclude in Section 7.
2. Questions in Our Questionnaire

Our questionnaire contains some questions. These questions were necessary to report the activities of our university. In this section, we show the questions in detail as follows:

1 [Experience through university life]
   How much did you emphasize your school life?
   - Lessons
   - Extracurricular activities
   - Part-time jobs
   - Volunteers
   - Seminars
   - Leaderships of research
   - Environments of research
   - Others

2 [Ability gained in our university]
   How much did you improve your ability?
   - Usability of English
   - Usability of languages except for English
   - Information processing
   - Unknown problems
   - Communication
   - Debating
   - Group activities
   - Professional fields
   - Analytics
   - Innovation
   - Reporting
   - Internationalization
   - Human and culture
   - Society

3 [Degree of achievement for education]
   How much did you achieve by education?
   - Liberal arts education
   - Professional education
   - Seminar
   - Graduation Research
   - Practice, internship, or volunteer
   - Leaderships of research
   - Environments of research
   - Others

4 [Degree of satisfaction for education]
   How much were you satisfied by education?
   - Liberal arts education
   - Professional education
   - Seminar
   - Graduation Research
   - Practice, internship, or volunteer
   - Leaderships of research
   - Environments of research

5 [Influence from teacher]
   Did you meet a teacher influenced you?
   - Meet a teacher motivating to study
   - Meet a teacher understanding troubles
   - Meet a teacher advising properly
   - Meet a teacher explaining well
   - Meet a teacher working hard
   - Meet a teacher impressing

6 [Time to study]
   How long did you spend time for study?
   - Lessons or experiments
   - Homework, preparations, reviews
   - Other studies
   - Talking with teachers
   - Extracurricular activities
   - Part-time jobs
   - Reading
   - Hobby
   - Volunteer

7 [Degree of usage for facilities]
   How much did you use facilities?
   - Library
   - Computer
   - Internet
   - Environment for Individual Study

8 [Degree of satisfaction for facilities]
   How much were you satisfied by facilities?
   - Library
   - Computer
   - Internet
   - Environment for Individual Study

9 [Degree of usage for school life support]
   How much did you use school life supports?
   - Syllabus
   - Guidance
   - Scholarship
   - Health care
   - Career path
   - Talk about harassment

10 [Degree of satisfaction for school life support]
    How much were you satisfied by school life supports?
    - Syllabus
    - Guidance
    - Scholarship
    - Health care
    - Career path
    - Talk about harassment

11 [Degree of recognition of university]
How much did you know your university?
- Educational charter
- Educational purpose
- Admission policy
- Curriculum policy
- Diploma policy

12 [Opinions to our university]
Please tell your opinions freely.

The number of questions is about 80. Most questions are five-alternative or two-alternative. In a few question, answerers can write freely.

3. Paper Questionnaire

In this section, we explain a paper questionnaire before explaining a web questionnaire. The paper questionnaire have some problems. We show some examples to show these problems. To solve them, we introduce a web questionnaire system we will mention in Section 4.

3.1 Example of Paper Questionnaire

We show three examples of the paper questionnaires: Hokkaido University, Mobility Management, and Hiroshima University. We show not only these examples but also these problems.

3.1.1 Hokkaido University

Hokkaido University conducted paper questionnaires that the targets were students and professors in 2006 [7]. The purposes of these questionnaires are to know the thoughts of students about the curriculum of this university, and the opinions about the education from the professors.

On the one hand the questionnaire for the students contained 15 questions, and on the other hand the questionnaire for the professors contained 9 questions. The response rates were 12.3% (333/2,717) and 48.6% (261/537) respectively. The authors described that the response rate of students was not high though the students received the questionnaire paper when registering their course.

In this questionnaire, the students could not earn anything after answering questions, and that they were not reminded to answer. So we thought that most of them lost their motivations to answer the questions.

3.1.2 Mobility Management

In general, the people don’t answer any questions though they are invited to a questionnaire. The one may forget to answer the questionnaire. The other may not notice this invitation. So reminding them to answer is very important.

Dillman described the Total Design Method (TDM) [8]. TDM is the method to increase response rate. The characteristics of TDM are as follows:

- Preparing regular patterns of invitation, reminder, and so on.
- Sending reminders to answerers many time.
- Decreasing the number of questions.

If the population of targets is clear, TDM is very efficient.

Hagihara et al conducted a questionnaire survey about Mobility Management (MM) by using paper [2]. MM is the measure to solve some traffic problems. They used reminder like TDM. They used reminders and rewards to increase the response rate. But the reminders and the rewards cost many expense.

3.1.3 Hiroshima University

In Hiroshima University, questionnaire surveys were conducted using mark sheet for the students to evaluate their lessons from 2004 to 2008 [9]. The questionnaire survey is held in each semester. The response rate was 72.7% (62,440/85,927) on average. The students received the questionnaire mark sheet at the end of the semester. So such a high response was obtained.

While there were many responses, many expenses was needed to conduct the paper questionnaire. The cost was 3,519,546 yen which was equivalent to about 35 thousand dollar. It was very expensive to convert the mark sheet data to the electronic data.

Moreover the authors said that there were weak points when conducting paper questionnaires. For example, the staff lost the paper which had been already answered, the anonymity of the answers was not kept, and so on.

4. Web Questionnaire

A paper questionnaire needs a lot of expense and labor as mentioned in previous section. So, in this section, to solve the problems occurred in paper questionnaire, we introduce a web questionnaire. After we show LimeSurvey [10] which is one of web questionnaire systems, we explain some works related to web questionnaire.

4.1 LimeSurvey

LimeSurvey is an open source Web questionnaire system.
This system has many sophisticated functions and we can use it free of charge. So we can reduce an initial cost to start a questionnaire survey. In our case, since we have own virtual environment, we also need no expenses to prepare
hardware. The version of this system we used is “1.92+ Build 120919”.

The system has 6 functions mainly.

1. System Management
2. Questionnaire Management
3. Design Template
4. Create Question
5. Mail
6. Total and Analysis

We mention the detail of these functions in the following.

4.1.1 System Management

We can create a user as a questionnaire manager and can give the user permissions to manage questionnaire. The user can give each of questionnaires permissions to execute, modify, delete, and read. Moreover, we can create groups the users belong to, and can give the groups some permissions.

The users created by system management functions can log in LimeSurvey. After logging in, they can operate some functions. There is a special user who can do everything just after installing LimeSurvey. The people who use LimeSurvey for the first time usually log in as this special user.

4.1.2 Questionnaire Management

The user who is given the permission to manage questionnaires can change a configuration of questionnaires. For example, the user can configure the day a questionnaire should start or end. A configuration of the questionnaire can be imported or downloaded by the user. Especially, to register answerers is one of the most important items. There are also some other items to manage questionnaires.

The user can manage multi questionnaires. The list of all questionnaires is available and the user can know the conditions of every questionnaires. Moreover, in each questionnaire, the user can create, delete, and modify some questions. Ordering questions function can be used. If there are questions the answerers have to answer, the user can assume the questions indispensable. In this way, the user can operate questionnaires or their questions variously.

A questionnaire can be replicated. We assume that the user would like to make a same questionnaire as the questionnaire having already conducted. The user can do such a replication easily after only the user downloads and uploads the configuration of the questionnaire.

4.1.3 Design Template

The user can change the design of a questionnaire page by selecting a design template. The design template is editable, importable, and exportable. Having the knowledge of HTML and the style sheet, the user can edit a design template easily.

4.1.4 Create Question

Some question forms are available. If the user would like to get an answer from a question, he or she can use single-selective question, such as radio button or pull-down menu. The user can adopt multi-selective question, such as checkboxes, to get multi answers. There are matrix-selective question, text box to describe freely, and so on.

The complex question forms are also available. The user can randomize the order of the choices. For example, the user also can let the answerer answer the question only if a condition is satisfied. Moreover, the created questions can be previewed.

When there is a question the user had already created, the user can replicate and reuse this question. If you would like to create many questions, you don't need to create all questions individually.

4.1.5 E-mail

The users can receive some notifications e-mails and can send some kind of e-mails to the answerers. In this section, intensively, we describe the e-mails to the answerers.

Before starting a questionnaire, the user can send an invitation e-mail to the answerers. This e-mail contains an explanation of the questionnaire, a URL linked to the questionnaire site, an e-mail address for inquiry, and so on. The URL contains a token which can specify a student. The answerers can start answering some questions after accessing the URL. After the answerers answer all of questions, the e-mails to thank for answering are sent to them.

The user can also send e-mails for reminding to the answerers who didn't finish answering all questions. The user can know who didn't complete to answer the questionnaire. Moreover they only send reminders to such answerers. The reminder e-mail also contains a URL linked to the questionnaire site. So, the answerers can start answering immediately after they receive it.

These e-mails are created based on each e-mail templates which the user prepared in advance. Using some specific methods to describe, the names of the answerers can be included in the e-mail. The e-mail function is very useful and contributed greatly to increase our survey response rate.
Moreover the user can receive the e-mail which notices something important. For example, the noticing e-mail is sent to the user when the introducing e-mail to the answerers failed to be sent.

4.1.6 Total and Analysis

The user can confirm the total of answers from questionnaire surveys. This system enables the user to show graphs created from the answers. If the user would like to operate the data of the answers, he or she can use CSV format files downloaded from the system. These functions are so useful to analyze the answers. By the results of these analysis, the user can decide or find something significant.

An environment of statistical computing “R” is used widely among statisticians. R is one of free software programing languages. LimeSurvey can create the file suitable for R. You can analyze many data statistically.

4.1.7 Others

As we mentioned above, LimeSurvey has many functions. In this section, we show the other functions to be explained especially.

When the answerers are answering a questionnaire, they know how long the questionnaire will have finished. LimeSurvey has progressing bar which tell the answerers what percent of total questions they have already answered. So the answerers feel easy.

Some of the answerers would like to stop answering and resume later. LimeSurvey also have a resumable function. Then they can restart answering at the point suspended.

4.2 Related Work

There are some works related to web questionnaire. First, we show an example using LimeSurvey as we also used. Next, we explain some reports about general web questionnaire. Finally, we describe some papers about problems and solutions of web questionnaire.

4.2.1 Questionnaire using LimeSurvey

In Ritsumeikan University, a paper questionnaire had been conducted until 2007 [11]. The staffs of this university knew that a web questionnaire was very efficient in terms of labor and expense. So they decided to conduct a web questionnaire survey using LimeSurvey in 2009.

They also knew some advantages and disadvantages when using web questionnaire systems. Major advantages as follows:

- No transportation expense
- Reduction of labor
- Traceability

These advantages are also important for us to conduct a questionnaire on Web. Especially, it is the most valuable advantage that the expense and the labor became low. As individual data are also stored in the web system, we can trace the answerers. Thus, web questionnaire has some advantages.

Next, major disadvantages as follows:

- Digital divide
- Low response rate
- System fault

The authors worried about the answerers who could not use Internet. Moreover, it is said that the response rate doesn’t become high when conducting web questionnaire. According to circumstance, the system may stop suddenly. Recently, there are few people who cannot use Internet. Especially, almost all university students are given an environment to use computers, mobiles, and so on. System fault hardly occur. So, the authors decided to use LimeSurvey.

LimeSurvey has a mailing function mentioned Section 3.1.5. However, the authors didn’t use this function because of a danger to individual information. Instead of this function, they sent e-mails to answerers manually. Also they sent postcards to remind them to answer the questionnaire. As a result, the response rate became 27.9% (326/1,168). They thought that the questionnaire played well.

Conducting this questionnaire, they needed some expense and labor despite the fact that they wanted to reduce them. Actually, sending postcards was very effective. We thought that this was equivalent to high expense they paid. It was better that they should have obtained high response rate with low expense.

4.2.2 General Web Questionnaire

In this section, we describe about general web questionnaire. Web questionnaires have some characteristics. Ohsumi et al classified web questionnaires by the characteristics of answerers [12] [13]. The classifications contain three types.

- [panel type] All people registered can answer.
- [open type] An indefinite number of people can answer.
- [resource type] Selected targets from all registered people can answer.

Moreover resource type is divided into three methods.
The questionnaire is conducted by applying open type for all registered people.

The targets satisfied certain conditions can answer.

Answerers are selected at random from the registered people.

In our university, the teachers applied for participating our questionnaire. The students belonging in such departments were targets. So we think that our questionnaire is close to resource type. Moreover, sampling method in the resource is the most similar to ours.

4.2.3 Problems

Although the web questionnaire is very useful, there are some problems. In this section, we show the example of the problems.

Kakiyama et al researched the quality of web questionnaire [14]. They said that the web questionnaire was easier and cheaper than the paper questionnaire. They also mentioned that they could not understand whether the answerers can answer exactly. For example, the identity of the answerer is not recognizable, that is to say, the other person may be able to answer. The answerers may not respond properly.

Moreover, Murase shows some problems using web questionnaire [15]. Generally, there are various answerers and they are partial in terms of age, gender, educational background, and so on. He said that the questionnaire result extracted from the random targets might be incorrect. If YES rate of a question is 60%, it may be 20% or 90% really. So, he insisted that it was important to pre-survey the age, the gender, and the educational background of the targets.

The targets of our web questionnaire were the students of our university. As we sent the students the own mail to introduce and remind, they could answer themselves. Moreover we had known their age, gender, and educational background. So, our environment of web questionnaire was properly.

4.2.4 Solutions

Many targets have not answered or don’t answer all questions in web questionnaire. This is one of the problems using web questionnaire. If these targets become to reduce, the response rate can be increased. There are the researchers to solve this problem.

Yamamori et al developed a web questionnaire system that the administrator could recognize who didn’t answer the questionnaire [16]. Moreover the answerers were anonymous. Generally, when a questionnaire is conducted, there are some requests as follows:

- High response rate
- One person can answer once
- Anonymity

The system that they developed satisfied these requests.

They used tokens to keep anonymity. A token is a character string to distinguish individuals. They sent the answerers the e-mail including the URL containing individual token. Each answerers could answer the questionnaire once. If an answerer completed to answer the questionnaire, an evidence was stored in the system. So they could recognize who could not complete to answer. Moreover the e-mails to remind could be sent to the answerers from the system, so the response rate was become to be increase.

This system was developed by using sympoll that is one of PHP program. Unfortunately, multi questionnaires could not conducted at the same time. If they realize this function, they had to operate the databases manually.

This solution becomes to be standard afterward. LimeSurvey, we used, also provided these functions. LimeSurvey has enough functions to increase response rate, to keep anonymity, to send reminders, and so on. Especially, reminder function is very useful. In next section, we mentioned the effect of reminders in detail.

5. How to remind

When we would like to gather more answers, reminding the answerers is very effective method. In this section, we compare a paper questionnaire and a web questionnaire from a point of reminding. Then we explain that reminders on web questionnaires are easier than those on paper questionnaires.

5.1 Reminder in paper questionnaire

It is difficult to remind the answerers to answer some questions in a questionnaire using papers. Especially, if the answerers cope with this questionnaire optionally, it is more difficult. Moreover the answerers tend not to answer any questions without rewards.

Hagihara et al researched the methods of questionnaires [2]. They carried out the paper questionnaires in three conditions as follows.

1. Handing over
2. Posting
3. Mailing
In their any methods, the cost they should pay was not free until the answerers can answer the questionnaire. They reminded the answerers to answer the questionnaire. Then, they also needed an expense to send some mails to remind.

Furthermore, they compared a questionnaire rewarding the answerers with a questionnaire not rewarding them. Naturally, the rate of those who were given some rewards was higher than those who were not given any rewards. The relation between the reminders and the rewards is shown as Table 1.

### Table 1

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Without rewards</th>
<th>With rewards</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without reminders</td>
<td>21.3%(17/80)</td>
<td>37.5%(30/80)</td>
</tr>
<tr>
<td>With reminders</td>
<td>28.8%(23/80)</td>
<td>57.5%(46/80)</td>
</tr>
</tbody>
</table>

In 2009, they started using MOMIJI. On the same year, they didn’t use reminder function, and then the response rate was 48.5%. On this time, the penalties were imposed, too. Next year, 2010, they started reminding the students to answer the questionnaire again and the response rate increased up to 76.0%. So, the reminders were very efficient to increase the response rate obviously.

While the reminders they sent worked very well, they did a troublesome task beforehand. Exactly, they had to analyze the data of answers to specify the students who should be sent an e-mail reminding to answer the questionnaire. In short, it was not easy to decide to send reminding mails for them.

Using Limesurvey which we introduced, we can know who don’t complete to answer the questionnaire. Moreover, we can send e-mails to them, after we click a few times only. In next section, we describe our web questionnaire provided by LimeSurvey.

### 5.2.2 LimeSurvey (Kyushu University)

We adopted a web questionnaire system, whose name was LimeSurvey, to obtain students' information. In the web questionnaire system, the user as an administrator of this system can remind the answerers to answer the questions of the questionnaire. To remind them, the user just clicks the button for reminding. After this action, e-mails to remind are sent to the answerers. The answerers receive the e-mail and can confirm that they must to answer the questionnaire.

We don’t need to prepare the papers of the questionnaire. Then we also need not to hand over, post, and mail. So there are no expense to remind. Moreover we don’t need to work for specifying the targets who should be sent an e-mail to remind.

We also can obtain other useful information from LimeSurvey. We can know how many the answerer had been sent the reminders. As we get this information, we can know the effectiveness of the reminders easily. For example, if an answerer didn’t access the web questionnaire system although the reminders were sent many times, it becomes clear that he or she may not read e-mails. In the case of our university, we don’t fail to send e-mails because all students have their own e-mail address given when they entered.

In our questionnaire, before we reminded at first, the response rate was 13.6%. We reminded the answerers three times. After reminding, finally the rate became 36.6%. The detail of this data will be shown in next section.
Table 2
The number of the answers in each period

<table>
<thead>
<tr>
<th>Period</th>
<th>Subtotal</th>
<th>Cumulative Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st period</td>
<td>371 (13.6%)</td>
<td>371 (13.6%)</td>
</tr>
<tr>
<td>2nd period</td>
<td>396 (14.5%)</td>
<td>767 (28.1%)</td>
</tr>
<tr>
<td>3rd period</td>
<td>140 (5.1%)</td>
<td>907 (33.2%)</td>
</tr>
<tr>
<td>4th period</td>
<td>92 (3.3%)</td>
<td>999 (36.6%)</td>
</tr>
</tbody>
</table>

6. Effect of reminding

In this section, we describe our questionnaire and its effect of reminding.

6.1 Abstract of our questionnaire

Our questionnaire survey was intended to hold from October 11 to November 11, 2013 in the original plan. The objects of the questionnaire survey were the students of Kyushu University. They were mainly final grade students in the graduate or undergraduate course. The number of students intended to answer the questionnaire was 2,730. We didn’t reward the students for answering. Also we didn’t impose any penalties on students. Moreover it was optional whether the answerers would answer the questionnaire.

At first, we sent invitation e-mails to the students on October 11, 2013. All students have own e-mail addresses which are given by our university. So, all invitation e-mails should have been sent to the students certainly.

We reminded the answerers three times to answer the questionnaire as following date.

- November 5, 2013
- November 14, 2013
- November 20, 2013

Additionally, we let the teachers, who led the students, remind them at the same time. As LimeSurvey can output the list of the students who didn’t complete to answer, we can give the teachers the list easily. On November 5, about a week before the end of the period, we sent reminders to them at first. Then we decided that we changed the end of the period to November 21 because the response rate was 13.6% which was too low as shown in Table 2. We sent second reminders to them on November 14, about a week before the new end of the period. Moreover we sent the last reminders to them on November 20 previous to the day the questionnaire survey ended.

6.2 Effect

We divide the period to answer the questionnaire into 4 periods as follows:

- 1st period: From October 11 to November 4, 2013
- 2nd period: From November 5 to November 13, 2013
- 3rd period: From November 14 to November 19, 2013
- 4th period: From November 20 on

Table 2 shows the number of the answers obtained in each period. Moreover Figure 1 shows the daily variation of the number of answers. In this graph, the vertical axis means the number of answers and the horizontal axis means the days. In Figure 1, 4 peaks are recognized. The first peak is found on the first day of the questionnaire, the others are found on the days we reminded the answerers to answer the questionnaire. We explain the effectiveness of the reminder in the following part.

At the first day the questionnaire started, 112 answers were obtained. But until the day we reminded the answerers to answer the questionnaire, the daily number of answers was decreasing. The subtotal of answers in the 1st period was 371, which was 13.6% in the whole.

To increase answers, we reminded the answerers at first on November 5, 2013. Then at the day the reminder had sent, 185 answers were obtained. The number of answers per day was the highest value. In the 2nd period, the subtotal was 396 (14.5%). Until the end of 2nd period, the cumulative total was 767 (28.1%). This result showed that the reminder was very effective to increase the number of answers.

We reminded the answerers again to increase the number of answers more and more until the questionnaire ended. So the second reminder was sent on November 14, 2013. As we expected, the answers on the day was 86, which was high value. In the 3rd period, the subtotal was 140 (5.1%) and the cumulative total was 907 (33.2%). The degree of increasing in 3rd period was lower than 2nd period, but we could confirm again the reminder was effective exactly.
The last reminder was sent on November 20, which was the day before the questionnaire was ended. Figure 1 shows that this reminder was also effective. After the day the questionnaire was planned to end, actually, we left this questionnaire system open. In the 4th period, the subtotal was 92 (3.4%).

Finally, the total of answers was 999 and the rate was 36.6%. If the reminders were not sent to the answerers, the rate increased up to 20.0% at most. So the reminders were very effective. Figure 2 shows the rate of answers in each period from within all answers. The 2nd period was the highest number of answers although the period is only 10 days. It was clear that the first reminder was most useful. 62.9% of all answers were obtained after the first reminder.

7. Conclusion

In this paper, we explained the effectiveness of the reminders in the web questionnaire survey. We introduced LimeSurvey as a web questionnaire survey system. Using this system, we could remind easily the answerers to answer the questionnaire. At the same time, the teachers of these students who were answerers worked to remind them to answer. Actually, we could increase the number of answers from our questionnaire survey using the reminder function. Some teachers also contributed to increase the number of answers. It is so important that the response rate was 36.6% although the answerers cope with this questionnaire survey optionally without any rewards or penalties.

In future work, we would like to examine the relation between the time of submitting answer and some questions. For example, we assume that there are students who answer the questionnaire early. If we find that the students study a long time, we find that the timing the students answer the questionnaire relate to the time of study.

There were many students who didn't answer the questionnaire. Some of them didn't read the invitation or the reminder e-mails. Other of them might not answer the questionnaire although they had read these e-mails. We will work hard that the response rate will increase using the web questionnaire survey.

References

[13] Osamu Yoshimura and Noboru Ohsumi,
“Quality of Internet Survey”, *ISM Symposium*, pp. 15-32, 2003


**Tetsuya Oishi** is an assistant professor in the office of institutional research at Kyushu University. He has the degree of “Doctor of Engineering”. Dr. Oishi worked as an engineer at NTT DATA KYUSHU Corporation from 2004, a Post-doctoral Fellow or a Technical Support Staff in Faculty of Information Science and Electrical Engineering at Kyushu University from 2009, and the present post from 2013. His fields of researches are data mining, information retrieval, institutional research, and so on.

**Eiichi Takata** is an associate professor in the office of institutional research at Kyushu University. He has the degree of “Master of Laws”. Mr. Takata worked at some places, for example a staff in Graduate School of Education and Human Development School of Education at Nagoya University, and the present post from 2007. His fields of researches are university evaluation, higher education, educational administration, institutional research, university management, and so on.

**Masao Mori** is an associate professor in the office of institutional research at Kyushu University. He has the degree of “Doctor of Information Science”. Dr. Mori worked as a Research Associate in Faculty of Information Science and Electrical Engineering at Kyushu University from 1996, an Associate Professor in the Office of Institutional Research at Kyushu University from 2006, and the present post from 2012. His fields of researches are data science, information retrieval, institutional research (IR), and so on.
Expression Strength for the Emotional Scene Detection from Lifelog Videos

Atsushi Morikuni\textsuperscript{a}  Hiroki Nomiya\textsuperscript{b}  Teruhisa Hochin\textsuperscript{c}
Kyoto Institute of Technology, Japan  Kyoto Institute of Technology, Japan  Kyoto Institute of Technology, Japan

Abstract

For the purpose of retrieving emotional scenes from a lifelog video database, and showing them, we propose a criterion to measure the strength of emotions. Conventionally, the retrieval of the emotional scenes has been performed by distinguishing the kind of the emotion of a person. However, precisely judging the importance of the scene is difficult without the consideration of the strength of the emotion. Therefore, we introduce a criterion called expression strength in order to measure the strength of emotions based on the amount of the change of several facial feature values. The effectiveness of the expression strength for the detection of the emotional scenes with smiles is shown through an experiment using a lifelog video data set.

Keywords: Lifelog, video retrieval, facial expression recognition, emotion.

1. Introduction

Due to the recent improvement of the performance of video cameras and storage devices, anyone can easily make a large quantity of multimedia data (e.g., the growth records of children). For such a reason, lifelog has attracted attention \cite{1}[2]. Lifelog is the recordings of everyday life and can be recorded as various types of data such as texts, images, and videos. Particularly, video data can be made easily and contain a wide variety of useful information. Therefore, we focus on lifelog videos in this study.

As recording videos becomes easy, an enormous amount of video data can be stored in the video databases. This makes the retrieval of lifelog videos quite difficult. Consequently, a considerable amount of valuable lifelog data is not utilized. Hence, an efficient and accurate retrieval of useful video scenes is indispensable to fully make use of lifelog videos.

For the better utilization of lifelog videos, an effective retrieval method has been proposed \cite{3}. This method detects impressive scenes from a lifelog video by using the facial expressions of a person in the video. It can retrieve emotional scenes that a person expresses a kind of facial expression. However, the intensity of the emotion (i.e., the intensity of the facial expression) cannot be estimated. For example, it can detect “smile” but it cannot distinguish “giggle” from “laughter.”

In order to improve the retrieval quality of emotional scenes, we introduce expression strength, which is the criterion to measure the strength of facial expression. The expression strength is calculated using several salient points on a face called facial feature points.

The effectiveness of the expression strength is evaluated by using a lifelog video data set and an image data set widely used for the evaluation of the quality of facial expression recognition. As a result, we reveal that the expression strength is able to estimate the intensity of facial expression and to find the scenes that the users want to retrieve.

The remainder of this paper is organized as follows. Section 2 presents some related works. Section 3 describes the facial features and the computation of the expression strength using the facial features. Section 4 shows the experiment to evaluate the usefulness of the expression strength. Finally, Section 5 concludes this study.

2. Related Works

2.1 Video Retrieval

Several video-sharing sites (e.g., YouTube \cite{4}) are widely used to retrieve videos from huge video databases. Many of them provide tag-based retrieval methods. In order to use the retrieval system, some users post their videos with several tags (or keywords) describing the contents of the videos and others search the videos by means of the tags given to the posted videos. This kind of retrieval system will be quite useful for publicly accessible videos.

On the other hand, the owner of lifelog videos must give the tags to each video one by one since most of lifelog videos are private ones. Despite the burden of the tagging
tasks, there are only a few scenes that the user wants to watch them again [5]. Hence, the tag-based retrieval is not suitable for lifelog video retrieval.

Content-based video retrieval methods provide more intuitive search frameworks [6]. They can retrieve various scenes by means of such as key frames (key images) on the basis of the similarity of some kinds of feature values obtained from the videos and the key frames. They do not require tagging tasks but it is difficult to retrieve useful scenes because selecting appropriate key frames is not easy.

2.2 Emotional Scene Detection Based on Facial Expression

In general, the users want to retrieve an interesting or impressive scene from lifelog videos. The person in such a scene will express a certain emotion (e.g., surprise, happiness, etc.). Therefore, finding the emotional scenes will be helpful to retrieve useful scenes. Because the emotion can be estimated from facial expressions, an emotional scene detection based on facial expression recognition has been proposed [3].

This method detects emotional scenes by means of the recognition of facial expressions in each frame image of a video. For each frame image, several facial features are computed from the positional relationships of a few facial feature points. Then, all the frame images in a video are classified into one of the predefined facial expressions. Finally, the emotional scenes are detected according to the classification result.

This method can retrieve emotional scenes without troublesome tasks such as tagging. However, it cannot estimate the intensity of the emotion. Estimating the intensity of emotion will be important because a scene with stronger emotion will be more interesting and impressive.

We thus introduce a criterion called expression strength for the estimation of the intensity of the emotion. Additionally, we conduct an experiment to estimate the strength of the emotions and verify the hypothesis that the users want to retrieve the scenes with strong emotions.

3. Expression Strength

The expression strength is a criterion to measure the strength of the emotion on the basis of the movement of salient points on a face. We call the points facial feature points. The movement of the facial feature points is represented by some facial features defined as the positional relationships of the facial feature points.

3.1 Facial Feature Points

As shown in Fig. 1, the set of facial feature points consist of 42 points on the following components of a face:

- Left and right eyebrows: 10 points ($p_1, \ldots, p_{10}$)
- Left and right eyes: 18 points ($p_{11}, \ldots, p_{28}$)
- A mouth: 14 points ($p_{29}, \ldots, p_{42}$)

These facial feature points are obtained by using an application software called FaceSDK 4.0 [7].

3.2 Facial Features

We define the following 11 types of facial features to estimate the expression strength. These facial features are based on the positional relationships of the facial feature points and are considered to be associated with the change in facial expressions.

1. Gradient of right and left eyebrows: $f_1$

This feature value is based on the gradients of the right and left eyebrows (denoted by $a_r$ and $a_l$, respectively) computed by using least squares. The gradients of the right and left eyebrows are obtained from the facial feature points $\{p_1, \ldots, p_3\}$ and $\{p_6, \ldots, p_{10}\}$, respectively. This feature value is defined as Equation (1).

$$f_1 = \frac{a_r - a_l}{2}$$  (1)

2. Distance between eyes and eyebrows: $f_2$

Using the average distance between the facial feature points on the eyebrows and the corresponding facial feature points on the upper side of eyes, the value of this feature is obtained through Equation (2).

$$f_2 = \frac{\sum_{i=1}^{10} ||p_{i} - \bar{p}_{i+10}||}{10 \cdot L}$$  (2)
Here, $L$ is a normalization factor for the difference of the size of a face. It is defined as the distance between the center points of left and right eyes, that is, $L = \|\vec{p}_{27} - \vec{p}_{28}\|$.

3. **Area between eyebrows:** $f_3$

This feature value is given by Equation (3) as the area of the quadrangle formed by connecting four facial feature points $p_5$, $p_6$, $p_{16}$, and $p_{15}$ located at the inner corners of eyebrows and eyes.

$$f_3 = \frac{S(p_5, p_6, p_{16}, p_{15})}{L^2} \quad (3)$$

Here, $S(p_1, \cdots, p_m)$ is the area of a polygon formed by connecting $m$ facial feature points $p_1, \cdots, p_m$.

4. **Area of eyes:** $f_4$

This facial feature is the normalized areas of two octagons formed by the facial feature points on the circumference of the left and right eyes, defined by Equation (4).

$$f_4 = \frac{1}{2L} \left\{ S(p_{11}, p_{12}, p_{13}, p_{14}, p_{15}, p_{25}, p_{22}, p_{21}) + S(p_{16}, p_{17}, p_{18}, p_{19}, p_{20}, p_{26}, p_{25}, p_{24}) \right\} \quad (4)$$

5. **Vertical-to-horizontal ratio of eyes:** $f_5$

Based on the ratio of the distance between the top and bottom points to the distance between the left and right points on the left and right eyes, this feature value is defined by Equation (5).

$$f_5 = \frac{1}{2} \left( \tan^{-1}\frac{\|\vec{p}_{22} - \vec{p}_{13}\|}{\|\vec{p}_{25} - \vec{p}_{18}\|} + \tan^{-1}\frac{\|\vec{p}_{25} - \vec{p}_{18}\|}{\|\vec{p}_{20} - \vec{p}_{16}\|} \right) \quad (5)$$

6. **Area of the circumference of a mouth:** $f_6$

This feature value is defined by Equation (6) as the normalized area of the octagon formed by connecting eight facial feature points located on the circumference of a mouth.

$$f_6 = \frac{S(p_{29}, p_{31}, p_{32}, p_{33}, p_{30}, p_{34}, p_{35}, p_{36})}{L^2} \quad (6)$$

7. **Area of inner circumference of a mouth:** $f_7$

Similar to the sixth feature value, this feature value is defined by Equation (7) as the normalized area of the octagon formed by connecting eight facial feature points located on the inner circumference of a mouth.

$$f_7 = \frac{S(p_{29}, p_{37}, p_{38}, p_{39}, p_{30}, p_{40}, p_{41}, p_{42})}{L^2} \quad (7)$$

The sixth feature value is influenced by the thickness of the lips which can vary depending on the type and the intensity of the facial expression. On the other hand, this feature value is hardly affected by the thickness of the lips.

8. **Vertical-to-horizontal ratio of the circumference of a mouth:** $f_8$

Based on the ratio of the distance between the top and bottom points to the distance between the left and right points on the circumference of a mouth, this feature value is defined by Equation (8).

$$f_8 = \tan^{-1}\frac{\|\vec{p}_{35} - \vec{p}_{32}\|}{\|\vec{p}_{30} - \vec{p}_{29}\|} \quad (8)$$

9. **Vertical-to-horizontal ratio of the inner circumference of a mouth:** $f_9$

Similar to the eighth feature value, this feature value is defined by Equation (9) based on the ratio of the distance between the top and bottom points to the distance between the left and right points on the inner circumference of a mouth. This feature value is also insensitive to the thickness of the lips.

$$f_9 = \tan^{-1}\frac{\|\vec{p}_{41} - \vec{p}_{38}\|}{\|\vec{p}_{30} - \vec{p}_{29}\|} \quad (9)$$

10. **Vertical position of the corner of a mouth:** $f_{10}$

This feature value represents how high the position of the corner of a mouth is. It is defined by Equation (10).

$$f_{10} = \frac{(y(p_{29}) + y(p_{26})) - (y(p_{35}) + y(p_{32}))}{y(p_{32}) + y(p_{35})} \quad (10)$$

Here, $y(p)$ is the $y$-coordinate of the facial feature point $p$. If the mean value of the $y$-coordinate of the facial feature points on the corner of a mouth is larger than that of the facial feature points on the top and bottom of a mouth, $f_{10}$ becomes positive. Thus, a larger value of $f_{10}$ represents a higher vertical position of the corner of a mouth.

11. **Angles of corners of a mouth:** $f_{11}$

This feature value is the average value of the angles of the left and right corners of a mouth. The angle of the left
(right) corner is formed by connecting the three facial feature points located on the left (right) corner of a mouth. It is given by Equation (11).

\[ f_{i1} = \frac{A(p_{29}, p_{31}, p_{36}) + A(p_{30}, p_{33}, p_{34})}{2} \quad (11) \]

where \( A \) is the function to compute the angle formed by three facial feature points \( p, q, \) and \( r \). \( A \) is defined by Equation (12).

\[ A(p, q, r) = \cos^{-1}\left( \frac{(\hat{p} - \hat{q}) \cdot (\hat{p} - \hat{r})}{\| \hat{p} - \hat{q} \| \| \hat{p} - \hat{r} \|} \right) \quad (12) \]

For each frame in a video, the above feature values are computed and the feature vector \( f_{i1}, \ldots, f_{i11} \) is obtained. Here, the \( j \)-th feature value obtained from the \( i \)-th frame from the beginning of the video is denoted by \( f_{ij} \).

### 3.3 Expression Strength

The expression strength is defined for a single frame taking into consideration of the tendency that the feature value is proportional to the strength of emotion. In order to accurately estimate the strength of emotion, the baseline of feature value is first determined using a training data set. The training data set consists of several emotional and nonemotional frames. An emotional frame is the frame that a person in the frame image expresses a certain emotion. On the other hand, a nonemotional frame is the frame that the facial expression of a person in the frame is neutral. The training set is manually prepared prior to the computation of the expression strength.

The training data set \( T \) is represented as \( T = \{ g_{1e} \cdots g_{N_e}, g_{1n} \cdots g_{N_n} \} \). Here, \( g_{ie} \) and \( g_{in} \) are the \( i \)-th emotional and nonemotional frames, respectively. \( N_e \) and \( N_n \) are the numbers of emotional and nonemotional frames in the training set, respectively. \( N_e \) and \( N_n \) have to be determined experimentally.

Because of the personal difference of the facial expressions, it is quite difficult to estimate the expression strength directly from the feature values. Hence, the baseline of feature values are computed to diminish the personal difference of the feature values.

The baseline feature value is determined for each facial feature. The baseline of the \( j \)-th facial feature \( S_j \) (\( j = 1, \ldots, 11 \)) is defined by Equation (13).

\[ S_j = \frac{\sum_{i=1}^{N_e} f_{ij} + \sum_{i=1}^{N_n} f_{nj}}{N_e + N_n} \quad (13) \]

Here, \( f_{ej} \) (\( f_{nj} \), respectively) is the \( j \)-th feature value of the \( i \)-th emotional (nonemotional) frame in \( T \).

The expression strength is computed on the basis of the difference between the feature values obtained from a frame and the baseline feature values. The expression strength of the \( i \)-th frame in a video \( E_i \) is defined by Equation (14).

\[ E_i = \sum_{j=1}^{11} (f_{ij} - S_j) \quad (14) \]

Therefore, the higher value of the expression strength represents the stronger emotion.

4. Experiment

4.1 Hypotheses

We make some hypotheses in regard to the usefulness of the expression strength for emotional scene retrieval. Through this experiment, we verify the following hypotheses:

1. The expression strength is able to represent the intensity of facial expressions.
2. The expression strength becomes high in the scenes that the users want to retrieve.
3. The expression strength is not affected by the movement of a mouth unrelated to the expression of emotions, which mainly appears in conversations.
4. The expression strength can be used for various persons.

Note that we focus on the emotion of happiness, which is one of the six basic emotions [8], because most of users will want to retrieve the scenes with happiness. The emotion of happiness often leads to the facial expression of smiles. Hence, we estimate the expression strength of smiles in this experiment.

4.2 Experimental Settings

1) Lifelog Videos

The experiment was conducted by eight subjects (termed Subjects A, B, C, D, E, F, G, and H), all of which were 21- to 28-year-old male university students. The subjects were divided into two disjoint groups (denoted by \( \alpha \) and \( \beta \)) consisting of four subjects. Subjects A, B, C, and D belong to \( \alpha \) and the other subjects belong to \( \beta \).

For each subject, the scenes of playing cards were recorded four times by using a web camera placed in front of the subject in order to record the subject's face. The \( i \)-th video of Subject \( X \) is denoted by \( X_i \) (\( i = 1, 2, 3, 4 \)). For
Subjects A, B, C, and D were asked to select the scenes they wanted to retrieve from the two videos recorded by the video camera. The videos recorded by the web cameras were used to show the videos to the subjects. The videos recorded by the video camera is shown in Table 1.

The videos recorded by the web cameras were used to estimate the expression strength of the subjects. Because of the high frame rate of these cameras, we selected frames from each video after every 10 frames in order to reduce the computational cost. For example, \( A_1 \) means the first video of Subject A. The specification of the web camera is shown in Table 1.

Apart from the web cameras, a video camera was used to record the scenes of playing cards including all the subjects in a group at the same time. Thus, a total of eight lifelog videos were recorded by the video camera. The \( i \)-th video for \( \alpha \) is denoted by \( \alpha_i \) (\( i = 1, 2, 3, 4 \)). Similarly, the \( i \)-th video for \( \beta \) is represented by \( \beta_i \). The specification of the video camera is shown in Table 1.

The videos recorded by the web cameras were used to estimate the expression strength of the subjects. Because of the high frame rate of these cameras, we selected frames from each video after every 10 frames in order to reduce the computational cost. For example, a total of 1800 frames are used for a 10-minute video.

On the other hand, the videos recorded by the video camera were used to show the videos to the subjects. The subjects were asked to select the scenes they wanted to retrieve from the two videos recorded by the video camera. Subjects A, B, C, and D were asked to select the scenes from \( \alpha_3 \) and \( \alpha_4 \). The other subjects were asked to select the scenes from \( \beta_3 \) and \( \beta_4 \). The scene was specified by the beginning time and the ending one in seconds. The length of each video is about ten minutes.

2) Learning Settings

Both of the parameters \( N_e \) and \( N_n \) in Equation (13) were experimentally set to 12, taking into consideration of the tradeoff between the accuracy and efficiency of the estimation of the expression strength.

Hence, the training set for a certain video consists of 12 emotional and 12 nonemotional frame images. They were randomly selected from all the frame images of the video from which FaceSDK accurately detected the facial feature points. It was determined by one of the authors whether the feature points were accurately detected or not. All the frame images other than the training images were used as the test set.

3) Strength of Smile

In order to evaluate the estimation accuracy of the expression strength, we first classify the smile into three levels according to the strength of the smile. We call the three levels of smile \( \text{Smile1} \), \( \text{Smile2} \), and \( \text{Smile3} \). The accuracy of the expression strength is evaluated by the average values of the estimated expression strength for the frames of \( \text{Smile1} \), \( \text{Smile2} \), and \( \text{Smile3} \).

The reason for this classification is that the correct expression strength must manually be determined to evaluate the estimation accuracy of the expression strength, but it is very difficult to manually determine the correct expression strength in real number. It seems to be reasonable that classifying the smile into three levels because many people will be able to discriminate the strength of a smile in only a few levels.

The definitions of \( \text{Smile1} \), \( \text{Smile2} \), \( \text{Smile3} \), and the other facial expressions are shown in Table 2. These are defined by one of the authors. He classified all the frames in the videos recorded by the web cameras into one of the expressions described in Table 2.

4.3 Experimental Result

1) Verification of the first hypothesis

We computed the expression strength for the videos \( A_1 \), \( B_2 \), \( C_3 \), \( D_4 \), \( E_1 \), \( F_1 \), \( G_1 \), and \( H_1 \) because most of the facial feature points were correctly detected in these videos. The average expression strength for each facial expression is shown in Table 3. Note that the frames other than \( \text{Smile1} \),

<table>
<thead>
<tr>
<th>Specifications</th>
<th>Web camera</th>
<th>Video camera</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>ELECOM UCAM-DLU130HWH</td>
<td>Sony HDR-CX560V</td>
</tr>
<tr>
<td>Frame size (in pixels)</td>
<td>640x480</td>
<td>1440x1080</td>
</tr>
<tr>
<td>Frame rate</td>
<td>30fps</td>
<td>30fps</td>
</tr>
<tr>
<td>Video format</td>
<td>WMV</td>
<td>MPEG4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Facial expression</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smile1</td>
<td>The subject pulls up the corners of his mouth and slightly opens his mouth.</td>
</tr>
<tr>
<td>Smile2</td>
<td>The subject lowers the corners of his eyes and smiles with his mouth open.</td>
</tr>
<tr>
<td>Smile3</td>
<td>The subject laughs with his mouth wide open.</td>
</tr>
<tr>
<td>Expressionless</td>
<td>The subject shows no expression.</td>
</tr>
<tr>
<td>Others</td>
<td>The subject shows a certain expression other than smile.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Video</th>
<th>Expressionless</th>
<th>Smile1</th>
<th>Smile2</th>
<th>Smile3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A_1 )</td>
<td>-0.163</td>
<td>0.136</td>
<td>0.166</td>
<td>0.258</td>
</tr>
<tr>
<td>( B_2 )</td>
<td>-0.345</td>
<td>-0.070</td>
<td>0.055</td>
<td>0.129</td>
</tr>
<tr>
<td>( C_3 )</td>
<td>-0.170</td>
<td>0.108</td>
<td>0.111</td>
<td>0.129</td>
</tr>
<tr>
<td>( D_4 )</td>
<td>-0.165</td>
<td>0.109</td>
<td>0.201</td>
<td>0.405</td>
</tr>
<tr>
<td>( E_1 )</td>
<td>-0.035</td>
<td>0.098</td>
<td>0.383</td>
<td>N/A</td>
</tr>
<tr>
<td>( F_1 )</td>
<td>-0.284</td>
<td>-0.361</td>
<td>0.187</td>
<td>0.364</td>
</tr>
<tr>
<td>( G_1 )</td>
<td>0.043</td>
<td>0.276</td>
<td>0.312</td>
<td>N/A</td>
</tr>
<tr>
<td>( H_1 )</td>
<td>-0.339</td>
<td>0.195</td>
<td>0.242</td>
<td>0.328</td>
</tr>
</tbody>
</table>
Table 4. Average expression strength for the scenes selected by each subject.

<table>
<thead>
<tr>
<th>Video</th>
<th>Sel.</th>
<th>Nonsel.</th>
<th>Expressionless</th>
<th>Smile1</th>
<th>Smile2</th>
<th>Smile3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₄</td>
<td>0.001</td>
<td>-0.129</td>
<td>-0.179</td>
<td>0.170</td>
<td>0.195</td>
<td>0.241</td>
</tr>
<tr>
<td>B₄</td>
<td>-0.033</td>
<td>-0.251</td>
<td>-0.405</td>
<td>-0.007</td>
<td>0.095</td>
<td>0.135</td>
</tr>
<tr>
<td>C₄</td>
<td>-0.006</td>
<td>-0.094</td>
<td>-0.210</td>
<td>0.023</td>
<td>0.131</td>
<td>0.148</td>
</tr>
<tr>
<td>E₄</td>
<td>0.231</td>
<td>0.003</td>
<td>-0.035</td>
<td>0.098</td>
<td>0.223</td>
<td>0.268</td>
</tr>
<tr>
<td>F₄</td>
<td>0.012</td>
<td>-0.025</td>
<td>-0.042</td>
<td>-0.002</td>
<td>0.107</td>
<td>0.214</td>
</tr>
<tr>
<td>G₄</td>
<td>0.210</td>
<td>0.067</td>
<td>-0.153</td>
<td>0.153</td>
<td>0.312</td>
<td>0.367</td>
</tr>
<tr>
<td>H₄</td>
<td>0.025</td>
<td>-0.222</td>
<td>-0.039</td>
<td>0.195</td>
<td>0.242</td>
<td>0.328</td>
</tr>
</tbody>
</table>

Table 5. Average expression strength for the scenes of conversations.

<table>
<thead>
<tr>
<th>Video</th>
<th>Conv.</th>
<th>Nonconv.</th>
<th>Expressionless</th>
<th>Smile1</th>
<th>Smile2</th>
<th>Smile3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>-0.163</td>
<td>-0.214</td>
<td>-0.039</td>
<td>0.136</td>
<td>0.166</td>
<td>0.258</td>
</tr>
<tr>
<td>B₂</td>
<td>-0.345</td>
<td>-0.468</td>
<td>-0.177</td>
<td>-0.070</td>
<td>0.055</td>
<td>0.129</td>
</tr>
<tr>
<td>C₂</td>
<td>-0.170</td>
<td>-0.358</td>
<td>-0.134</td>
<td>0.108</td>
<td>0.111</td>
<td>0.129</td>
</tr>
<tr>
<td>D₂</td>
<td>-0.165</td>
<td>-0.356</td>
<td>-0.028</td>
<td>0.109</td>
<td>0.201</td>
<td>0.405</td>
</tr>
<tr>
<td>E₁</td>
<td>-0.035</td>
<td>-0.120</td>
<td>0.050</td>
<td>0.098</td>
<td>0.383</td>
<td>N/A</td>
</tr>
<tr>
<td>F₁</td>
<td>-0.284</td>
<td>-0.385</td>
<td>-0.164</td>
<td>-0.361</td>
<td>0.187</td>
<td>0.364</td>
</tr>
<tr>
<td>G₁</td>
<td>0.043</td>
<td>-0.013</td>
<td>0.042</td>
<td>0.276</td>
<td>0.312</td>
<td>N/A</td>
</tr>
<tr>
<td>H₁</td>
<td>-0.339</td>
<td>-0.645</td>
<td>-0.136</td>
<td>0.195</td>
<td>0.242</td>
<td>0.328</td>
</tr>
</tbody>
</table>

Smile2, Smile3, and Expressionless are excluded and that Subjects E and G have no frame of Smile3.

For most of the subjects, the average expression strength for each facial expression has the relationship that “Expressionless < Smile1 < Smile2 < Smile3.” This result indicates that the expression strength can appropriately represent the intensity of smile.

The expression strength of Smile1 is less than that of Expressionless in regard to Subject F. The subject's facial expression of Smile1 is very similar to that of Expressionless, compared with the other subjects. Although the accuracy of the estimation of the expression strength can slightly be affected by the personal difference, we can verify the first hypothesis to some extent from this experimental result.

2) Verification of the second hypothesis

We compared the estimated expression strength for each facial expression with that in the scenes selected by the subjects. The videos used in this experiment are A₄, B₄, C₄, E₄, F₄, G₄, and H₄ because the facial feature points are accurately detected in these videos. The video of Subject D was not used because the facial feature points were not accurately detected from both D₃ and D₄ (Note that the subjects selected the scenes only from α₃, α₄, β₃, and β₄).

The average expression strength is shown in Table 4. The columns “Sel.” and “Nonsel.” represent the average expression strength obtained from the frames selected by the subjects and the other frames, respectively.

For all the subjects, the average expression strength of the selected scenes is higher than that of the other scenes. There are a total of 22 scenes selected by all the subjects. 20 scenes out of the 22 scenes include the frames of Smile2 and/or Smile3. From the fact that the expression strength for Smile2 and Smile3 is considerably high for all the subjects, the second hypothesis can be verified.

3) Verification of the third hypothesis

Since some of the facial features are associated with the facial feature points on a mouth, the conversations between the subjects may affect the expression strength. In order to clarify the influence of the conversations on the expression strength, we computed the expression strength of the scene of conversations using the same videos as used in the first experiment.

The average expression strength is shown in Table 5. The columns “Conv.” and “Nonconv.” represent the expression strength of the conversation scenes and non-conversation scenes, respectively. Note that both scenes do not include any of Smile1, Smile2, and Smile3.

The average expression strength of the conversation scene is much less than that of Smile1 for most of subjects. Although the average expression strength of the conversation is higher than that of Smile1 for Subject F, it is much lower compared with the expression strength of Smile2. From this experimental result, the conversation has very little influence on the expression strength and the third hypothesis can be verified.
4) Verification of the fourth hypothesis

For the verification of the effectiveness of the expression strength for various persons, we used the Cohn-Kanade AU-Coded Facial Expression Database [9], which is widely used for the evaluation of the facial expression techniques. This data set contains the sequences of facial images of 18- to 30-year-old males and females. We used the image sequences of 18 subjects in this experiment. The image sequence starts with a neutral face and ends with a full smile (i.e., the first image is a neutral face and the last image shows a full smile) as shown in Fig. 2. Hence, the intensity of the smile is proportional to the image number. The number of images in an image sequence depends on the subject and varies from 12 to 33. The average number of images in an image sequence is about 20.9.

Because the number of images in an image sequence is small, it is impossible to obtain sufficient training images from this data set. Hence, in this experiment, we define the expression strength for a video (denoted by $\varepsilon$) while the expression strength is defined for a frame in the above experiments. The value of $\varepsilon$ is defined by Equation (15).

$$\varepsilon = \sum_{j=1}^{m} \left( \sum_{i=1}^{11} f_{i(M-1)} - \sum_{j=1}^{11} f_{i,j} \right)$$

(15)

Here, $M$ is the number of images in the image sequence. This value is defined on the basis of the difference of the feature values between the first $m$ images and the last $m$ images in the image sequence. We set the value of the parameter $m$ to 3 considering the average length of the image sequence. The expression strength (i.e., the value of $\varepsilon$ of each subject) is shown in Fig. 3.

For all the subjects, the expression strength is positive. From this result, it can be detected by the expression strength that the intensity of the facial expression in the images of a smiling face is higher than that in the images of a neutral face. Although there are personal differences of the expression strength, smiles can be detected from all the subjects by means of the expression strength. Therefore, the fourth hypothesis can be verified from this experimental result.

5. Conclusion

For the purpose of an accurate retrieval of the emotional scenes in lifelog videos, we propose a criterion called expression strength in order to measure the intensity of the emotion. The expression strength is efficiently computed by the geometric computation in regard to the facial feature points. The experimental result showed that the expression strength can appropriately estimate the strength of smiles for various persons.

It is, however, unclear whether the expression strength is effective for diverse facial expressions as well as smiles. Verifying the usefulness of the expression strength for a wide variety of facial expressions is included in the future work.

The emotional scene retrieval method using the expression strength has not yet been developed. Developing the lifelog video retrieval system by introducing an efficient emotional scene retrieval algorithm is also one of the future works.

References


facial expression recognition based on machine learning. He has given presentations at national and international conferences and workshops. He is an active member of the IEEE Computer Society and the Information Processing Society of Japan and the Japanese Society for Artificial Intelligence. He has reviewed papers for the Institute of Electronics, Information and Communication Engineers (IEICE) Transactions on Information and Systems.

Teruhisa Hochin is a professor in the department of Information Science at Kyoto Institute of Technology. His research interests include graph-based, multimedia, scientific, and flexible databases, extensibility of DBMS, and Kansei retrieval. He is a member of IEEE, ACM, Institute of Electronics, Information and Communication Engineers (IEICE), Information Processing Society of Japan (IPSJ), Database Society of Japan (DBSJ), Japan Society for Archeological Information (JSAI), and Japan Society of Kansei Engineering (JSKE) as well as ACIS. He has given presentations at national and international conferences, symposiums, and workshops. He has reviewed papers for ACIS International Journal of Computer & Information Science, ACIS International Journal of Software Innovation, IEICE Transactions on Information and Systems, IPSJ Journal, IPSJ Transaction on Databases, DBSJ Journal, Journal of JSIAI, Kansei Engineering International Journal, Transactions of JSKE, and Journal of Communication and Computer.

Atsushi Morikuni is received his ME degree in Information Science from Kyoto Institute of Technology. His research interests include multimedia data engineering and information retrieval.

Hiroki Nomiy is an assistant professor in the department of Information Science at Kyoto Institute of Technology. His research focus is multimedia data engineering, image recognition and machine learning. He has been engaged in the research of multimedia data retrieval, especially emotional video scene retrieval using
An Effective and Appealing Instruction Based on Active Learning in a Large Class of Project-based Research Learning on Environmental Problems

Susumu Yamazaki\textsuperscript{a}, Tsukasa Morimoto\textsuperscript{a}, Tohru Futawatari\textsuperscript{a}, Atsushi Nogami\textsuperscript{a}
University of Kitakyushu, Japan

Taku Jiromaru\textsuperscript{b}
OME, Inc., Japan

Masafumi Iwano\textsuperscript{c}
NEOZEST, Japan

Tokuro Matsuo\textsuperscript{d}
Advanced Institute of Industrial Technology, Japan

Abstract
We apply workshop-based instruction on project-based research learning on environmental problems for approximately 280 first-year undergraduate students. Our instructional design is to read the manuals and record activities and issues; to categorize them according to the eight essential points of project-based learning; and to share the results with teams nearby. Our assessment results show that this workshop is effective for team project-based research learning; that satisfaction with the workshop is high; that the workshop is effective in reducing the anxiety; and the workshop encouraged each student to participate in the group work.

Keywords: Active learning, collaborative learning, cooperative learning, project-based learning, problem-based learning, research learning, workshop, large class, environmental education.

1. Introduction

Kitakyushu City has much experience in solving serious pollution problems, and has been designated an Eco-Model City (EMC) by the Japanese government. The University of Kitakyushu has also been teaching about the environment as pictured in the fig.1, and established Faculty of Environmental Engineering in 2001 to train students as environmental engineers. We have conducted or been involved in project-based research learning, called Case Studies of Environmental Issues, which teaches upwards of 250 first-year undergraduate students in Faculty of Environmental Engineering. This class has the following two problems:

1. The 25 tutors vary in terms of skill and motivation because they have not been trained.

2. Instruction for students has been ineffective because they are bored by its lecture-style format.

Thus, we have tried to instruct tutors and students in the use of a workshop style of instruction in this course. Originally, this course is active learning because it is designed as project-based learning, which is a kind of active learning. However, the instruction, which is placed at begging of the course, is not active-learning-style format but lecture-style format. Thus, we also apply active learning approaches to the instruction.

In this paper, we introduce instructional design of the workshop, and assess its effectiveness and appeal.

The organization of the rest of this paper is as follows: Section 2 shows related works of this presentation including active learning and some related words, and case studies on environmental problems. Section 3 describes our previous course. Section 4 shows the instructional design of the workshop. Section 5 evaluates the effectiveness and appeal of the workshop and the course. Finally, Section 6 concludes this paper and previews our future work.

This paper is an extended version of work presented in the conference ESKM 2014 [1].

2. Related Work

2.1 Active Learning

Active learning is popularized in 1990s, triggered by Bonwell and Eison [2]. The definition of active learning is varied because different authors in the field have interpreted it and some related words differently [3]. Center for
Research on Learning and Teaching (CRLT) defines that active learning is a process whereby students engage in activities, such as reading, writing, discussion, or problem solving that promote analysis, synthesis, and evaluation of class content [4]. We show definitions of some words related to active learning:

- **Collaborative learning** is any instructional method in which students work together in small groups toward a common goal [3].
- **Cooperative learning** is a structured form of group work where students pursue common goals while being assessed individually [3].
- **Problem-based learning** is an instructional method where relevant problems are introduced at the beginning of the instruction cycle and used to provide the context and motivation for the learning that follows [3].
- **Project-based learning** is a model that organizes learning around projects [5].

Many authors have discussed instruction methods, strategies and evaluation methods of active learning. Sherman’s work [6] is one of them compiled in 1990’s.

### 2.2 Case Studies of Environmental Education

Because there are many case studies of environmental education, it is very hard to conduct a thorough and complete survey on it. Even though, we will show a part of it in this section.

- Brown and Palincsar [7] have conducted environmental education in elementary school. They focus reciprocal teaching and active listening. Their reciprocal teaching combines expert scaffolding, guided practice in applying simple concrete strategies, which consist of questioning, clarifying, summarizing and predicting, and cooperative learning discussion.
- Åhlberg et al. [8] provide environmental education using Knowledge Forum for schoolteachers and graduate students.
- Dahlgren and Öberg [10] illuminate how students raises questions about environment in the learning process by environmental education. They categorize the questions into five types: encyclopaedic, meaning-oriented, relational, value-oriented and solution-oriented.
- Stauffacher et al. [12] have conducted the transdisciplinary case study based on self-regulated learning and project-based learning. Self-regulated learning is learning that a learner is motivated by him/herself, adopts a strategy by him/herself, and understand in a metacognitive way.
- Klopfer and Squire [13] have conducted environmental education using handheld applications named augmented reality educational gaming.

### 3. The course “Case Studies of Environmental Issues”

University of Kitakyushu provides the course “Case Studies of Environmental Issues”. Fig. 1 shows a part of the curriculum tree directly related to the course. An arrow means a prerequisite relationship. For example, the course “Environmental Issue Special Lecture” is a prerequisite of the course “Case Studies of Environment”.

<table>
<thead>
<tr>
<th>Environmental Issue Special lecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st grade 1st semester</td>
</tr>
<tr>
<td>Omnibus lectures: great as special lecture in the environment person in charge</td>
</tr>
<tr>
<td>Tour of inspection: Eco-town, Kitakyushu Museum of Natural History, &amp; Human and so on</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case Studies of Environmental Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st grade 2nd semester</td>
</tr>
<tr>
<td>Field work for community</td>
</tr>
<tr>
<td>For the selected theme, mixed teams of the department make a concrete proposal and problem-solving</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Engineering Ethic</th>
</tr>
</thead>
<tbody>
<tr>
<td>2nd grade 1st semester</td>
</tr>
<tr>
<td>Case Study: learning organization’s role and social responsibility</td>
</tr>
<tr>
<td>Proposing a solution ourselves for each case</td>
</tr>
</tbody>
</table>

**Fig. 1 A Part of the Curriculum Tree**

The best way to understand the nature of the environmental problems is composed of the following steps:
1. Contact directly sites that are always faced with the environmental problems;
2. Identify key factors from various elements; and
3. Discuss why the problems have arisen.

In this course, the students will do as follows:
1. Organize teams;
2. Choose research topic in each team;
3. Identify the problems related to the topic;
4. Clarify the core of the problems;
5. Plan the goal of each team;
6. Select research methods suitable for the problems;
7. Share the tasks of each team;
8. Conduct research in self-regulation;
9. Summarize the research results; and
10. Make a presentation of them.

Thus, the students conduct both research and project on environmental problems. This course is a kind of active learning and project-based learning.

The research topics are provided by the instructors. The number of them is 25. The students are organized into 25 teams. Each team has approximately 10 students and a tutor.

- Week 6: Make an interim presentation;
- Week 7 to 8: Conduct research;
- Week 9 to 10: Summarize the research and prepare the preliminary presentation;
- Week 11: Make a preliminary presentation;
- Week 12-13: Summarize the research and prepare the final report and presentation;
- Week 14: Make a final presentation; and
- Week 15: Be praised for the final presentation.

Fig. 2 shows the flowchart of the schedules.

This paper focuses on the instruction at the week 1. This instruction was provided in lecture-style format, and the students had a complaints and dissatisfaction on it.

4. Instructional Design

Fortunately, we have instruction manuals for student participation in this course. We analyzed them and found out that this course has two features:

1. This course concerns research learning on an environmental problem; and
2. It is a team project.

We list the features of the abovementioned approach, and call these the eight essential points of project-based research learning:

1. Research learning:
   - Web-based surveys;
   - Bibliographic surveys;
   - Interviews; and
   - Presentations.

2. Project:
   - Team building;
   - The Plan-Do-Check-Act (PDCA) cycle;
   - Team communication: reporting, contacting and consulting; and
   - Reflection.

We then decided to create an instruction guide by listing activities recommended for this course and issues for each activity.

The eight essential points are functioning as advance organizers [14], which are information that is presented prior to learning and that can be used by the learner to organize and interpret new incoming information [15], or analogies useful to learning, including Analogical Thinking-Based Learning, Instruction Method based on Analogical Projection, or Instruction using Analogical Dropping [16-18].

![Flowchart of the Course Schedules](image-url)
4.1 Instructional Objectives

According to Gagne’s method [19], we define the instructional objectives of this guide as follows: Students will

- O1: state the eight essential points of project-based research learning (verbal information);
- O2: state the activities that should be conducted in this course and issues for each activity (verbal information);
- O3: classify items listed in O2 according to the eight points of O1 (intellectual skills, defined concepts); and
- O4: choose to cooperate as a team to conduct a project-based study (attitude).

These are multiple integrated instructional objectives or enterprises [19] to achieve the final objective of “conducting project-based research.”

4.2 Instructional Strategies

Following the instructional objectives, we define instructional strategies of this guide as follows:

1. For O4, we adopt team workshop-based instruction. This is in accordance with “attention” of the ARCS model [20].
2. For O2, the students will read the manuals and list activities and issues on sticky notes.
3. For O1, the students will prepare two papers on the eight essential points of project-based learning.
4. According to O3, the students will categorize each sticky note in step 2 into the eight points in step 3.
5. For greater effectiveness, the students will share the results of step 4 with nearby teams.

4.3 Implementation

It is not easy to conduct a workshop-based class for upwards of 250 students all at once. Thus, we first conducted four training sessions for the 25 tutors to evaluate our instructional design. We used four sessions to instruct the tutors in smaller groups and to evaluate and refine our instructional design in an iterative manner. Furthermore, we asked the tutors to write a critique. These critiques suggest that our design has their strong support. They identify some slight problems that we can solve immediately. Thus, we have gained confidence in our design.

Approximately 280 students attended the course. In the 180 minutes available, we conducted the session as follows (see Figure 2):

1. Icebreaker:
   A) Instruct the students to gather around in teams.
   B) Instruct them to wear a sticky note with their nicknames written on them as a name tag.
   C) Instruct them to introduce themselves to one another in 10 minutes, providing the following information:
      i. The origin of their nickname.
      ii. Hometown and alma mater.
      iii. Interest in this class.
2. Explain facilitation briefly as follows:
   What can be facilitated? (Meetings, projects, organizations, classes, civic activities, etc.)
3. Instruct them to practice basic brainstorming with sticky notes and a piece of paper:
   A) Instruct them to read the manual on procedures and issues in facilitation.
   B) Instruct them to brainstorm “how to make a class enjoyable”:
      i. They have five minutes to write ideas; and
      ii. 25 minutes to share them with their team members.
4. Explain the eight essential points of project-based research learning:
   A) Explain each research item briefly, because the students are already familiar with them.
   B) Explain each project item in more detail, because the students are not familiar with them.
5. Take a 10-minute break, and instruct the team leader to allocate reading instruction manuals.
6. Instruct them to follow the instructional strategies in section IV.
7. Instruct them to ask the instructor questions and report any issues causing anxiety by e-mail after the workshop.
8. Provide them all with the answers to question raised in step 7.

5. Assessment

After the course, we surveyed the 280 students. The main questionnaire items were as follows:

- Q1: Did you feel anxiety before this course? (Very much, a little, not very much, not at all)
- Q2: What caused you anxiety before this course? (Team buildings, the PDCA cycle, team communication, reflection, Web-based surveys, bibliographic surveys, interviews, presentations, another aspect, none)
- Q3: What caused you anxiety before this course? (Open question)
Q4: Did you contribute to your team? (Very much, a little, not very much, not at all)

Q5: Did your team perform well? (Very well, reasonably well, not very well, not well at all)

Q6: Which task did your team do well? (Team buildings, the PDCA cycle, team communication, reflection, Web-based surveys, bibliographic surveys, interviews, presentations, another aspect, none)

Q7: Specifically, what did your team do well? (Open question)

Q8: Which task did your team do badly? (Team buildings, the PDCA cycle, team communication, reflection, Web-based surveys, bibliographic surveys, interviews, presentations, another aspect, none)

Q9: Specifically what did your team do badly? (Open question)

Q10: Do you have any ideas as to why your team was unsuccessful in these areas? (Open question)

Q11: Did you understand the overview of this course? (Very much, a little, not very much, not at all)

Q12: Was the instruction provided in the guide of this course useful for understanding of the overview? (Very much, a little, not very much, not at all, I forgot the guide)

Q13: Describe what were you glad to learn by the instruction. (Open question)

Q14: Describe any suggestions to improve the instruction provided in the guide. (Open question)

Q15: Describe what were you glad to learn through this course. (Open question)

Q16: Describe any suggestions to improve this course. (Open question)

We received 76 valid responses. From this survey, we found:
1. This workshop is effective in usefulness for understanding the overview of the course.
Sixty-one reported that the instruction of the tutorial was useful (Q12, See Fig. 3). In particular, 34 students answered that it was nice to understand the research process (Q13, See Fig. 4).

2. Satisfaction with the workshop was high because this workshop helped each student understand the course process and become conscious of his/her role in the team. Seventy students answered their team perform well (Q5,
See Fig. 5) and Sixty-one reported the instruction of the tutorial was useful (Q12, See Fig. 3).

3. **The workshop is effective in reducing the anxiety.** Sixty-two of the 76 students felt anxiety before the workshop (Q1, See the left side of Fig.6). Fifty-seven (and 70 of the total of 76 respondents) answered that their teams performed well (Q5, See the right side of Fig.6). In particular, 49 students reported anxiety about team building and 31 reported anxiety about presentation before the workshop (Q2, See the left side of Fig.7), but performed well thereafter (Q6, See the right side of Fig.7). The workshop is effective in reducing the anxiety that students feel about group work and the process of this course.

4. **The workshop encouraged each student to participate in the group work,** and 55 students answered that they had experience of solving problems through group work (Q15).

5. Some problems on this course are suggested in open questions.
   - Regarding improvements to the instruction provided, 20 students answered that explanations should be brief. This does not mean that this workshop should be brief, because not only workshop-based instruction was provided, but also lecture-based instruction. As proof, some of them wrote the name of another instructor who was not present at this workshop in the open question. Still, it seems clear that in future, we should reduce the length of lecture-based instruction and explanations in the workshop.
   - Some students answered that their team performed badly, and described some ideas for improvement, for example the course schedule and time allocation, approaches to differences in members' motivation and the difficulty of each subject. This suggests that some students in the team may have less motivation to conduct research, which may prevent group work. Having identified this problem, in future, we will introduce some measures to address this problem.

6. **Conclusion**

In this study, we conducted workshop-based instruction for approximately 280 students of project-based research learning on environmental problems based on the following instructional design:

1. Read the instruction manual on the course and write activities and issues on sticky notes.
2. Classify each sticky note into the eight essential points of project-based learning (Team buildings, the PDCA cycle, team communication, reflection, Web-based surveys, bibliographic surveys, interviews and presentations).
3. Share the result of this work with nearby teams.

We conducted a survey of students after this course, and found the following:
1. This workshop is effective in usefulness for understanding the overview of the course.
2. Satisfaction with the workshop was high because this workshop helped each student understand the course process and become conscious of his/her role in the team.
3. The workshop is effective in reducing student anxiety about group work and the process of this course.
4. This workshop encouraged each student to participate in the group work, and many students reported that they had experienced problem solving through group work.

We conclude that this workshop was an effective way to conduct team project-based research learning. In future, we will reduce the length of lecture-based instruction and explanations in the workshop and will introduce some measures to address the problem of less motivated students. We will also propose application of this method to transform instruction, which tends to bore students, into attractive active learning.

Acknowledgements

We would like to thank Hiroki Shimono, Hiroyuki Tsujii, Nobuyuki Tachi and all tutors and students for discussions about the development and implementation of this workshop.

References


Susumu Yamazaki is a lecturer in the faculty of Environmental Engineering at University of Kitakyushu. Dr. Yamazaki's research focus is software engineering, marketing, and instructional design. His mission is to train undergraduate and graduate students as software engineers with entrepreneurship and instructional designers. He organized some study groups on these areas in Japan.

Tsukasa Morimoto is a professor in the Center for Fundamental Education, Hibikino Campus at the University of Kitakyushu. Prof. Morimoto’s research specialty is the relationship between philosophy of life and problem-solving thinking. His classes at the university are ‘Case Studies of Environmental Issues’, ‘Basic Ways of Thinking’, ‘Introduction to Ethics’, etc. His interests are human's attitude in the evolutionary epistemology, relation between evolutionary epistemology and philosophy of life, and especially ethics as procedures of solving problems (or coping with problems). See http://www.kitakyu-u.ac.jp/env/lang_en/faculty/center/tsukasa_morimoto.html

Tohru Futawatari received M.S. degree from School of Engineering, Kyushu University in 1984, and D.E. degree from Kyushu University in 1993. He had been a research associate University of Kitakyushu from 1984 to 1995. Then he moved to the University of Occupational and Environmental Health, to be an associate professor from 1995 to 2004. Then, he moved to the University of Kitakyushu, and is a professor in the Faculty of Environmental Engineering and Graduate School of Environmental Engineering since 2003. His research interests are mainly focus on Regional environmental management, Environmental management system, environmental accounting, reporting and labeling, environmental risk and risk management, environmental management for social sector and NGO/NPO, corporate social responsibility.

Atsushi Nogami is a Professor of Environmental Engineering since 2002, University of Kitakyushu. Current research and educational interest is in the area of environmental engineering. He was a chief researcher for computational material science at the R&D laboratory of Nippon Steel Corporation where he had worked from 1984 to 2002. The degree of Doctor of Engineering was awarded by University of Tokyo in 1984 after graduate study from 1979 to 1984 at the Department of Nuclear Engineering.

Taku Jiromaru is a Ph.D. candidate of Graduate School of Science and Engineering, Yamagata University. He also founded and served CEO of OME Inc. and ConferenceService Inc, since 2009 and 2011, respectively. He is a member of JSET, CIEC, JPCATS, and several other academic societies. His current research interest areas include Educational Technology, Mathematical Education and Career Service System.

Masafumi Iwano is a data analyst and web designer. Mr. Masafumi has experience of the error management, marketing and error analysis of the mobile prototype device. Now, Mr. Masafumi works on web designing, web direction, and the management planning analysis.

Dr. Tokuro Matsuo is a professor at Advanced Institute of Industrial Technology since 2012. He received the Doctor degree of Engineering from Dept. of Computer Science at Nagoya Institute of Technology in 2006. He is a research fellow of SEITI in Central Michigan University since 2010; was a research project
professor at Nagoya Institute of Technology between 2011 and 2014; was a research fellow at Shanghai University between 2010 and 2013; and a chair of director of International Institute of Applied Informatics since 2010. He was a visiting researcher at University of California at Irvine in 2010-2011. His current research interests include electronic commerce and business, service science and marketing, business management, artificial intelligence, material informatics, tourism informatics, convention research, incentive design on e-services. Some of his researches are presented in international conferences on AAAI, IEEE CEC, AAMAS, and WWW. He gave over 50 keynotes and invited talk at international conferences, symposia, and seminars. He also received over 40 awards and research grants from research foundations, company and Government. He is also commissioned as Adviser of Information Promotion from Japan Government.
Presumption Model for Postoperative Hospital Days from Operation Records

Takanori Yamashita\textsuperscript{a,c}  \hspace{1cm}  Yasuhiro Nakashima\textsuperscript{b}  \hspace{1cm}  Naoki Nakashima\textsuperscript{a}  
Kyushu University, JAPAN  \hspace{1cm}  Kyushu University, JAPAN  \hspace{1cm}  Kyushu University, JAPAN  
Yoshifumi Wakata\textsuperscript{a}  \hspace{1cm}  Yukihide Iwamoto\textsuperscript{a}  \hspace{1cm}  Sachio Hirokawa\textsuperscript{a}  
Kyushu University, JAPAN  \hspace{1cm}  Kyushu University, JAPAN  \hspace{1cm}  Kyushu University, JAPAN  
Satoshi Hamai\textsuperscript{a}  \hspace{1cm}  Brendan Flanagan\textsuperscript{a}  
Kyushu University, JAPAN  \hspace{1cm}  Kyushu University, JAPAN  

Abstract

The digitalization of medical treatment has progressed, and the importance of medical data utilization is increasing. Electronic medical data include structured digital data and unstructured text data. The secondary use of clinical text data to improve the quality and the efficiency of medical care is gaining much attention. However, there are few previous researches that have given feedback to clinical situations. The present paper analyzes the words that appear in operation records to predict the postoperative length of stay. SVM (support vector machine) and feature selection are applied to predict if a stay is longer than the standard length of 25 days. It was confirmed that with less than 20 feature words we can predict if a stay is longer or not with almost the optimal prediction performance, and evaluate specific words where prediction performance was low.

Keywords: Critical pathway, Operation record, SVM (support vector machine), Feature selection.

1. Introduction

1.1 The Secondary Use of Clinical Text Data

The digitalization of treatment information has progressed, and a vast quantity of clinical data has accumulated in hospital information systems. It is expected that the secondary use of this data will enhance medical quality, increase efficiency, and promote the resolution of medical social issues.

At the government and academic society level there are several projects that carry out data collection from two or more medical institutions with the aim of constructing a database and practical use, such as the Medical information database infrastructure improvement project [1], and the National Clinical Database [2]. The first project collects prescription and test result data with the aim of detecting medicinal side effects. The second project cooperates with a certified surgery medical system. Many of these projects that are aiming to structured data.

In Kyushu University Hospital, the data stored in the hospital information system is used for the secondary purpose of improving business-like statistics, clinical study, and nursing. Most of this is structured numerical data. Reports and summaries, in which essential consideration of medical examinations are written in free text as unstructured data, are not fully utilized.

There are some previous text mining researches on clinical data that analyzed hospital discharge summaries and reports. In [3, 4, 5] the extraction of feature words and the visualization of word relations were considered. The estimation of the name of a disease or a cancer classification was studied in [6, 7, 8, 9]. The usefulness in the analysis of the treatment program for each patient was reported in [10, 11], [12, 13] studied risk analysis. In the biomedical field, many researches that use text mining for analysis of a gene and protein are also seen in [14, 15, 16]. Furthermore, health professionals and researchers expect text mining as a supportive tool for decision-making and classification of the clinical process and the preventive care [17, 18].

There are some case studies that analyze medical examination text data, however, there are not many that are directly feedback to clinical practice or hospital management.

There are two reasons that make it difficult to provide feedback. The first is that the secondary use of text data is technically difficult, and the second is that many processes and laborious work are required [19, 20]. The present paper analyzes surgery operation records based on the clinical critical pathway.
1.2 Critical Pathways

The critical pathway in medical treatment is the same concept as an engineering critical pathway. It is a schedule of the course of medical treatment and recovery processes, such as examination and surgery, which is considered to be optimal for an inpatient. Since the therapeutic process and the contents are visualized, it is also easy for the patient or family to understand. For medical personnel, medical standardization is advanced, with roles and schedules becoming clear, making it possible to offer high quality medical treatment. The improvement effect of the medical management by offering medical treatment efficiently is also anticipated. It has been introduced positively in recent years in many medical institutes.

Standardization beyond the walls of medical institutions is required, and with the implementation of standard practice guidelines, it is anticipated that analysis result comparisons can be done between medical institutions. The Japanese Society of Clinical Pathway [21] mainly promotes the study of optimal critical pathways.

The critical pathway only assumes the optimum process at each time. Therefore, PDCA activity is always required for analysis, evaluation, and improvement. Although many analyses for an improvement have been reported, most of them use only structured data. In the past, as far as the authors know, there is almost no research using non-structured data, such as free description text data.

2. Research Purpose

In [22, 23, 24], we reported the analysis models with the postoperative length of stay as an objective variable, and patients’ attributes and critical pathway outcomes as an explaining variable. A critical pathway outcome is a component of the critical pathway that determines the goal of each step of the treatment. On the other hand, there are examples that consider outcome creation from log data and nursing order that is not included in the critical pathway [25, 26].

The purpose of this research is the development of the secondary use of medical-examination text data. A mechanism is built to extract the feature words that appear in the operation record of the patient who is hospitalized for a long period of time when compared with the standard duration of hospitalization. It is thought that those feature words are useful for the improvement of the medical-examination process and medical management. A method introduced in [27], which uses SVM (support vector machine) with feature selection, is applied to the operation records of hip replacement arthroplasty. Characteristic words are extracted to predict if the postoperative length of a patient is longer than the expected standard length (25 days). The effect of a medicine dictionary is evaluated by factor analysis.

3. Evaluation of Critical Pathway Outcomes

3.1 Structure of Critical Pathways

Figure 1 displays the structure of the clinical critical pathway, which consists of:
1) the outcomes as performance goals,
2) the course of medical treatment assumed to be the optimal at present (tasks), and
3) the observation items (assessments).
It is expressed as a schedule from hospital admission to discharge [28]. There are two kinds of outcomes: daily outcomes, and the final outcomes. The length of stay, termination, and destination after leaving hospital, medical expenses, and patient satisfaction are the final outcomes.

All variance type outcome pathways are the process of medical treatment that keeps all the records of clinical service and process in time series. Each record describes whether a task is performed or not and if an assessment or an outcome is achieved or not.

![Figure 1. The 3 Layers of Outcome, Assessment and Task](image)

3.2 Employment and Analysis of All Variance type Outcome Pathways

Kyushu University Hospital introduced the all variance type outcome critical pathway in 2008 [28]. The system forms a three-layered structure that connects an outcome with the assessment and the task (Figure 1).
A variance occurs when an outcome performance goal is not attained by evaluation in an assessment layer. Doctors and nurses keep records of the variance not only for laboratory data but for outcome as well.

4. Operation Records of Hip Replacement Arthroplasty

An operative record is a record of procedure, progress and observations of the target region during operation mentioned by the doctor.
4.1 Data for Analysis

The present paper analyzed 871 operation records of the hip replacement arthroplasty carried out in the period of January 2008 to March 2014 in Kyushu University Hospital. As the target variable of analysis, we prepared the postoperative length of stay obtained from the patients’ hospital discharge day and the operation day. However, in this experiment, the data of the patient ID that specifies a patient individual, sex, and a birth date are unnecessary, and were excluded from analytical data.

4.2 Postoperative Length of Stay

Postoperative hospital days are one of the last outcomes, and are a very important index for a patient and for a medical institution. Since the optimal hospital days influence a hospital bed occupancy rate directly in hospital management, they serve as a very important index. Moreover, it also has an influence on a patient's own cost directly.

Several years ago, the postoperative length of stay for the pathway of the hip replacement arthroplasty in Kyushu University Hospital was 27 days, and is now set to 25 days.

4.3 Characteristic Words Extraction by Search Engine

In order to analyze the 871 operation records, we constructed a search engine of the records. We used GETA [29] system available at NII. Using this search engine, we tried extraction of the word that may serve as a determinant of postoperative hospital days.

The number of words that appeared in all the reports was 2144. We used a medical dictionary that contains more than 40,000 words. However, only 406 words appear in the reports of the analysis. There are 282 dictionary words that appear more than once. If we include non-medical ordinary words, there are 1,318 words.

5. Prediction Reference of Postoperative Length Stay by SVM

5.1 Feature Words of Operation Records for Long Postoperative Stay

We indexed medicine dictionary words that appear in the target operation records and constructed a search engine. We used this search engine to apply SVM (support vector machine) to predict if the postoperative length of stay is longer than or equal to 25 days, which is the standard length in the current critical pathway of Kyushu University Hospital. We tried feature selection to obtain characteristic words that describe the operation records of longer stays.

The specific procedure is as follows. All operation records are vectorized using medicine dictionary words. If the length of the postoperative stay of a patient is longer than or equal to 25 days, then the record of the patient is marked as a positive data. Otherwise it is marked as negative data. Then the classification model is constructed based on the linear kernel by SVM. The score(wi) of each word wi is obtained by applying the model to the imaginary document that contains only the word. In [27], the score(wi) was used for the feature selection. In the present paper, we propose another two measures to evaluate the importance of each word. The first measure df(wi)*score(wi) is obtained as the product of the document frequency df(wi) of the word and the score(wi). The second measure log(df(wi))*score(wi) is the product of the log of the document frequency of the word and the score(wi).

The top 10 positive and negative words for the three measures are shown in Table 1, 2 and 3. The words in Table 1 that are chosen by score(wi) have small document frequencies. In Table 2, in which the words are chosen by df(wi)*score(wi), there are many words with large document frequencies. The words in Table 3 are selected by log(df(wi))*score(wi). There are a variety of document frequencies in Table 3.

<table>
<thead>
<tr>
<th>Word</th>
<th>score</th>
<th>df*sc</th>
<th>log(df)*sc</th>
<th>df</th>
<th>Word</th>
<th>score</th>
<th>df*sc</th>
<th>log(df)*sc</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>extrophy</td>
<td>7.82</td>
<td>93.78</td>
<td>19.42</td>
<td>12</td>
<td>cranial</td>
<td>-13.93</td>
<td>-125.39</td>
<td>-30.61</td>
<td>9</td>
</tr>
<tr>
<td>huge</td>
<td>7.45</td>
<td>52.12</td>
<td>14.49</td>
<td>7</td>
<td>washer</td>
<td>-13.87</td>
<td>-97.09</td>
<td>-26.99</td>
<td>7</td>
</tr>
<tr>
<td>tonic</td>
<td>6.83</td>
<td>20.48</td>
<td>7.50</td>
<td>3</td>
<td>slight</td>
<td>-11.67</td>
<td>-851.75</td>
<td>-50.06</td>
<td>73</td>
</tr>
<tr>
<td>varus</td>
<td>5.82</td>
<td>64.00</td>
<td>13.95</td>
<td>11</td>
<td>CT</td>
<td>-10.51</td>
<td>-42.04</td>
<td>-14.57</td>
<td>4</td>
</tr>
<tr>
<td>fracture line</td>
<td>4.76</td>
<td>33.29</td>
<td>9.26</td>
<td>7</td>
<td>anteroinferior</td>
<td>-10.34</td>
<td>-72.38</td>
<td>-20.12</td>
<td>7</td>
</tr>
<tr>
<td>scar</td>
<td>4.41</td>
<td>273.47</td>
<td>18.20</td>
<td>62</td>
<td>joint cavity</td>
<td>-9.64</td>
<td>-134.91</td>
<td>-25.43</td>
<td>14</td>
</tr>
<tr>
<td>sharp</td>
<td>4.33</td>
<td>99.56</td>
<td>13.57</td>
<td>23</td>
<td>TM</td>
<td>-8.58</td>
<td>-68.63</td>
<td>-17.84</td>
<td>8</td>
</tr>
<tr>
<td>artificial joints</td>
<td>4.31</td>
<td>64.66</td>
<td>11.67</td>
<td>15</td>
<td>MM</td>
<td>-7.98</td>
<td>-31.90</td>
<td>-11.06</td>
<td>4</td>
</tr>
<tr>
<td>diaphysis</td>
<td>4.14</td>
<td>45.56</td>
<td>9.93</td>
<td>11</td>
<td>hematoma</td>
<td>-7.97</td>
<td>-31.88</td>
<td>-11.05</td>
<td>4</td>
</tr>
<tr>
<td>leg position</td>
<td>3.88</td>
<td>11.65</td>
<td>4.26</td>
<td>3</td>
<td>central</td>
<td>-7.80</td>
<td>-23.40</td>
<td>-8.57</td>
<td>3</td>
</tr>
</tbody>
</table>
Table 2. Top 10 Feature words (by df*score)

<table>
<thead>
<tr>
<th>Word</th>
<th>score</th>
<th>df*sc</th>
<th>log(df)*sc</th>
<th>df</th>
<th>Word</th>
<th>score</th>
<th>df*sc</th>
<th>log(df)*sc</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>part</td>
<td>1.56</td>
<td>1140.24</td>
<td>10.27</td>
<td>732</td>
<td>slight</td>
<td>-11.67</td>
<td>-851.75</td>
<td>-50.06</td>
<td>73</td>
</tr>
<tr>
<td>proximal</td>
<td>1.92</td>
<td>1046.52</td>
<td>12.08</td>
<td>546</td>
<td>lateral</td>
<td>-4.81</td>
<td>-759.63</td>
<td>-24.34</td>
<td>158</td>
</tr>
<tr>
<td>load</td>
<td>2.45</td>
<td>932.95</td>
<td>14.55</td>
<td>381</td>
<td>adduction</td>
<td>-1.05</td>
<td>-541.16</td>
<td>-6.56</td>
<td>515</td>
</tr>
<tr>
<td>trial</td>
<td>1.78</td>
<td>884.71</td>
<td>11.05</td>
<td>497</td>
<td>dissection</td>
<td>-3.09</td>
<td>-535.07</td>
<td>-15.94</td>
<td>173</td>
</tr>
<tr>
<td>reposition</td>
<td>1.55</td>
<td>849.52</td>
<td>9.76</td>
<td>549</td>
<td>dislocation</td>
<td>-0.74</td>
<td>-454.78</td>
<td>-4.76</td>
<td>613</td>
</tr>
<tr>
<td>neck</td>
<td>1.84</td>
<td>824.40</td>
<td>11.25</td>
<td>447</td>
<td>external rotation</td>
<td>-0.56</td>
<td>-386.00</td>
<td>-3.63</td>
<td>695</td>
</tr>
<tr>
<td>cartilage</td>
<td>2.53</td>
<td>777.29</td>
<td>14.50</td>
<td>307</td>
<td>internal rotation</td>
<td>-0.27</td>
<td>-199.87</td>
<td>-1.79</td>
<td>737</td>
</tr>
<tr>
<td>X</td>
<td>1.45</td>
<td>720.35</td>
<td>9.00</td>
<td>497</td>
<td>fix</td>
<td>-0.36</td>
<td>-166.87</td>
<td>-2.22</td>
<td>462</td>
</tr>
<tr>
<td>X-rays</td>
<td>1.44</td>
<td>710.50</td>
<td>8.95</td>
<td>492</td>
<td>joint cavity</td>
<td>-9.64</td>
<td>-134.91</td>
<td>-25.43</td>
<td>14</td>
</tr>
<tr>
<td>piriformis muscle</td>
<td>2.04</td>
<td>698.35</td>
<td>11.89</td>
<td>343</td>
<td>cranial</td>
<td>-13.93</td>
<td>-125.39</td>
<td>-30.61</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 3. Top 10 Feature words (sort by log(df)*score)

<table>
<thead>
<tr>
<th>Word</th>
<th>score</th>
<th>df*sc</th>
<th>log(df)*sc</th>
<th>df</th>
<th>Word</th>
<th>score</th>
<th>df*sc</th>
<th>log(df)*sc</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>extrophy</td>
<td>7.82</td>
<td>93.78</td>
<td>19.42</td>
<td>12</td>
<td>slight</td>
<td>-11.67</td>
<td>-851.75</td>
<td>-50.06</td>
<td>73</td>
</tr>
<tr>
<td>load</td>
<td>2.45</td>
<td>932.95</td>
<td>14.55</td>
<td>381</td>
<td>washer</td>
<td>-13.87</td>
<td>-97.09</td>
<td>-26.99</td>
<td>7</td>
</tr>
<tr>
<td>cartilage</td>
<td>2.53</td>
<td>777.29</td>
<td>14.50</td>
<td>307</td>
<td>joint cavity</td>
<td>-9.64</td>
<td>-134.91</td>
<td>-25.43</td>
<td>14</td>
</tr>
<tr>
<td>huge</td>
<td>7.45</td>
<td>52.12</td>
<td>14.49</td>
<td>7</td>
<td>lateral</td>
<td>-4.81</td>
<td>-759.63</td>
<td>-24.34</td>
<td>158</td>
</tr>
<tr>
<td>varus</td>
<td>5.82</td>
<td>64.00</td>
<td>13.95</td>
<td>11</td>
<td>anteroinferior</td>
<td>-10.34</td>
<td>-72.38</td>
<td>-20.12</td>
<td>7</td>
</tr>
<tr>
<td>sharp</td>
<td>4.33</td>
<td>99.56</td>
<td>13.57</td>
<td>23</td>
<td>TM</td>
<td>-8.58</td>
<td>-68.63</td>
<td>-17.84</td>
<td>8</td>
</tr>
<tr>
<td>lateral position</td>
<td>2.39</td>
<td>522.64</td>
<td>12.86</td>
<td>219</td>
<td>marrow</td>
<td>-7.75</td>
<td>-69.74</td>
<td>-17.03</td>
<td>9</td>
</tr>
<tr>
<td>proximal</td>
<td>1.92</td>
<td>1046.52</td>
<td>12.08</td>
<td>546</td>
<td>dissection</td>
<td>-3.09</td>
<td>-535.07</td>
<td>-15.94</td>
<td>173</td>
</tr>
<tr>
<td>piriformis muscle</td>
<td>2.04</td>
<td>698.35</td>
<td>11.89</td>
<td>343</td>
<td>gluteus minimus</td>
<td>-7.69</td>
<td>-53.83</td>
<td>-14.96</td>
<td>7</td>
</tr>
</tbody>
</table>

5.2 Prediction Performance by Feature Selection

The top N positive words and the top N negative words are used to construct a model to predict if the postoperative length of stay is longer than or equal to 25 days. We evaluated the prediction performance by precision, recall, F-measure and accuracy. We used 5-fold cross validation in the evaluation experiment.

The accuracy of the baseline which uses all words is 0.4767. As seen in Figure 2, the best performance according to the score is obtained at N=30, whose accuracy is 0.6889. The other measure df*score and log(df)*score yield peak accuracies with small N, of 0.6580 at N=5 and 0.6206 at N=4 respectively. The peak accuracies with larger N are 0.6572 at N=60 for df*score and 0.7000 at N=50 for log(df)*score.

The F-measure of the baseline is 0.4872. As seen in Figure 3, the measure by score yields the best F-measure 0.7798 at N=30. The measure by df*score attains two peaks of 0.7680 at N=5 and 0.7648 at N=60. The measure by log(df)*score has two peaks of 0.7030 at N=4 and 0.8018 at N=50.

The prediction performance by score has a peak at N=50 with respect to both accuracy and F-measure. It is worthwhile to note that both df*score and log(df)*score give almost the same level of prediction performance with less than 10 words.
5.3 Effect of Dictionary Words

We conducted similar experiments based on ordinary words without using the medicine dictionary. Figure 4 and Figure 5 show the prediction performance of accuracy and F-measure obtained by the top N words. The plots of accuracy and F-measure with feature selection by score (red line) and df*score (green line) in Figure 4 and Figure 5 look similar to those in Figure 2 and Figure 3. The feature selection by df*score are effective compared to that by score. The feature selection by log(df)*score is effective with respect to the dictionary words as we can see that the blue lines have a steep increase in Figure 2 and Figure 4. However, the blue lines in Figure 4 and Figure 5 look similar to the red lines. Thus, the effect of feature selection by log(df)*score(wi) is not effective for ordinary words.
5.4 Investigation of the Ineffective Sections of Feature Selection by Ordinary Words

As seen in Figure 4 and 5, the accuracy and F-measure of Log(df)*score(wi) is extremely low at N=7. The precision and recall of ordinary words is also shown in Figure 6 and 7.

To find more, we focused on the recall of log(df)*score(wi) (blue lines) as it shows the same tendency as the accuracy and F-measure as seen in Figure 7. A positive word with a large document frequency appears for the first time at N=5. The positive word is “load” (document frequency: 385) and the recall improves as a result of its appearance as a general feature word.
However, at $N=7$ recall is lower than 0.2. It is thought that the decline in recall is influenced by the negative word “attach” which has a comparatively large document frequency (document frequency: 118). Therefore “attach” is not a negative word that is representative of postoperative length of stays that are longer than or equal to 25 days.

Figure 6. Precision by Ordinary Words

Figure 7. Recall by Ordinary Words
6. Conclusion and Future Work

Operation records were analyzed to obtain feature words that characterize long-term hospitalization. Three measures were proposed and the prediction performance was evaluated with respect to each measure. The first measure (score(wi)) was the one introduced in [27] and obtained by evaluating the score of each word against an SVM model of all words. The second measure (df(wi)*score(wi)) is defined as the product of the document frequency of the word and the score. The third measure (log(df(wi))*score(wi)) is defined as the product of the log of the document frequency of the word and the score of the word. When targeting medical dictionary words, the feature selection by df*score and by log(df)*score with less than 10 words resulted in the best prediction performance that is similar to the baseline score measure. Thus, these words can be considered as the characteristic words that describe long-term hospitalization. Moreover, we found that it is possible to evaluate specific words by reviewing the precision and recall around a portion of the feature selection of ordinary words where prediction performance was low.

The feature word was extracted regardless of the date of the operations in this experiment. We considered 25 days as the standard length of postoperative stay. However, in field practice, as the medical-examination process is improved from time to time, the critical pathway can also be corrected accordingly. The change of the standard length of stay may affect the distribution of the postoperative length of stay. Thus, we cannot guarantee that the feature words we obtained in the present paper characterize the patients who stay long-term. These words might represent a change of clinical procedures in the hospital. To clarify the two distinctions is a crucial future task we have to consider. Indeed, the analysis of temporal overlap relations of critical indicators are mentioned and left as further research in [22, 23, 24].

By incorporating the dates as features of operations, we plan to conduct research to distinguish the feature words of particular dates and that of particular patients. We aim to establish a text-mining technique that can perform clinical evaluation.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 24500176

References


---

**Yasuharu Nakashima** is an Associate Professor in the Department of Orthopedic Surgery, Graduate School of Medical Sciences at Kyushu University. His main research interests are in orthopedic biomaterials and kinematic analyses of hip joint.

**Yukihide Iwamoto** is a Professor in the Department of Orthopedic Surgery, Graduate School of Medical Sciences at Kyushu University. His main research interests are in basic and clinical research for malignant bone and soft tissue tumor.

**Brendan Flanagan** is a Doctorate student studying advanced information technology at the Graduate School of Information Science and Electrical Engineering, Kyushu University. His main research interests are in text mining, search engines, and text mining to support language learning.

**Naoki Nakashima** is a Professor in the Medical Information Center at Kyushu University Hospital. His main research interests are in process analysis of chronic diseases and in cost-effectiveness in medical service.
Takanori Yamashita is a Technical Specialist in the Medical Information Center at Kyushu University Hospital, and is a Doctorate student studying advanced information technology at the Graduate School of Information Science and Electrical Engineering, Kyushu University. His main research interests are in text mining, data mining and medical informatics.

Yoshifumi Wakata is an Assistant Professor in the Medical Information Center at Kyushu University Hospital. His main research interests are in process analysis of medical practice and quality of health care.

Sachio Hirokawa is a Professor in the Research Institute for Information Technology at Kyushu University. His main research interests are in text mining and search engine.

Satoshi Hamai is an Assistant Professor in the Department of Artificial Joints and Biomaterials, Graduate School of Medical Sciences at Kyushu University. His main research interests are in kinematic analyses of hip and knee joints.
Energy Efficient and Enhanced-type Data-centric Network Architecture

Shanming Zhang¹, Hidetoshi Takeshita, Satoru Okamoto, Naoaki Yamanaka
Keio University, Japan

Abstract

Information-centric networking (ICN) as an alternative has been researched for future Internet architecture. In this paper, a novel network architecture called Energy Efficient and Enhanced-type Data-centric Network (E³-DCN) is proposed for ICN. E³-DCN not only realizes ICN but also has the ability of data generation by organizing and combining relevant resource such as original data and kinds of processing services. In addition, in order to improve energy efficiency of data transmission, E³-DCN chooses optimal transmission path to transmit data based on packet switching and optical switching. Moreover, based on current traffic situation, E³-DCN reconfigures network topology such as unused nodes and links to operate in sleep mode for energy saving dynamically. And meanwhile, data caching strategies are applied into E³-DCN for data distribution based on the distribution of users. Data caching shortens transmission distance, so data transmission energy, latency and network load is decreased. Based on these approaches, E³-DCN achieves an energy efficient network.

Keywords: ICN, Future Internet, E³-DCN, Data Generation, Energy Efficient, Data Caching

1. Introduction

Current Internet architecture, which is host-centric communication model based on Transmission Control Protocol/Internet Protocol (TCP/IP), focuses on location where data exists. User must request specified server to get desired data using Uniform Resource Locator (URL) or IP address. Along with the increasing of Internet users and services, there are some fatal issues emerged e.g., network congestion, location dependence, mobility, network power consumption, Quality of Service (QoS), so that the current Internet cannot meet future Internet’s requirements [1]. Therefore, Information-centric networking (ICN) has been proposed for future Internet. ICN, which changes traditional communication model from “host-to-host” to “user-to-data”, focuses on what data is contained in network [2]. The basic idea of ICN is that user doesn’t consider where data exists and directly use data name to request and get data. Moreover, ICN also supports multiple communication ways such as unicast, multicast and so on. ICN has been explored by several projects such as Content-centric Networking (CCN) [3], Networking of Information (NetInf) [4] [5], Data-centric Networking (DCN) [6] [7].

In this paper, we propose a novel network architecture called Energy Efficient and Enhanced-type Data-centric Network (E³-DCN) for ICN. E³-DCN not only realizes ICN but also realizes a self-organizing network that can achieve optimization of resources and network energy consumption. E³-DCN has the ability of data generation. It can organize and combine relevant resource, such as original data and kinds of processing software and services as data materials, to generate desired data. In E³-DCN, users directly use data name to request and get desired data. When the requested data is not found, E³-DCN automatically organizes and combines relevant data materials to generate the requested data. In addition, E³-DCN also takes into account the network energy consumption issue. In order to improve energy efficiency of data transmission, E³-DCN chooses optimal data transmission path and transmission technology to transmit data based on packet switching and optical path/circuit switching. Moreover, according to current traffic situation, E³-DCN can dynamically reconfigure network topology for saving power consumption of the network such as unused data links and switches are shut down. Advanced data caching strategies, which make E³-DCN like energy efficient Content Delivery Network (CDN), are used for data distribution. Because data caching makes the distance of data transmission become shorter, so data transmission energy, latency and network load can be decreased. And meanwhile, data caching makes E³-DCN better support mobility of mobile devices which can get data from the nearest data source quickly while moving.

A prototype of the E³-DCN system is designed and constructed on Japan Gigabit Network-eXtreme (JGN-X) which is a multiple layer network testbed of Japan and can provide slice networks by network virtualization technology [8] (Detail is introduced in next section). E³-DCN is composed of two overlay networks on three slice networks. Two overlay networks, one is Data Centric Overlay Network (DCON) realized ICN, and another is Data Generation Overlay Network (DGON) used for data generation. Three slice networks are Control Plane Slice (CPS), Packet Switching Slice (PSS) and Circuit Switching Slice (CSS). Three slice networks as network infrastructure of E³-DCN are used for data transmission and network energy efficiency optimization.

The rest of the paper is organized as follows: In section two, research background and related work, including ICN research projects, Ubiquitous Grid Networking Environment

¹School of Science for Open and Environmental Systems, Graduate School of Science and Technology, Keio University.
3-14-1 Hiyoshi, Kouhoku-ku, Yokohama, 223-8522, Japan
shanming.zhang@yamanaka.ics.keio.ac.jp
(uGrid) and JGN-X, are described. In section three and four, we present our proposal E³-DCN architecture and energy optimized strategies in detail. In section five, the related evaluations are presented. In section six, we describe our experiments and E³-DCN prototype base on JGN-X. At last, we give our conclusions and future work.

2. Background

2.1 ICN Research Projects

CCN is an entirely new architecture based on name-based routing for ICN. In CCN, user sends “Interest” packet with data name to request desired data. “Data” packet as response is returned from data source following the track of “Interest” packet. And meanwhile, the transmitted data is cached at every intermediate node for future requests. CCN realizes the concept of ICN, but there are some problems existed in CCN such as every node holds a large number of data names as routing information cause routing delay problem. And it is very difficult to manage a huge number of cached data [9]. NetInf, which is also proposed for ICN, adopts hybrid name-based routing and Name Resolution Server (NRS) that implements matching among data name and IP address. User directly sends request holding data name to get data. The request is firstly forwarded toward NRSs based on name-based routing. If an NRS is available, data name can be resolved into a set of IP addresses of servers holding desired data. Subsequently, NRS chooses the best server to request data based on IP-based routing [4][5]. Since NetInf adopts name-based routing and IP routing, NetInf can easily connect different networks into a single Information-centric network. Therefore, NetInf has better scalability than CCN. However, delay problem still exists due to a large number of data names as routing information. Furthermore, NRS optimal distribution is also an important issue [10][11]. DCN is also proposed for ICN. DCN meets not only needs of human communication but also needs of Machine-to-Machine (M2M) communication for future Internet [12][13][14]. DCN adopts “route attraction” and “route aggregation” to reduce routing information comparing to CCN and NetInf. Furthermore, it makes the management of data cached become easier than CCN and NetInf [7]. Nevertheless, DCN doesn't take into account network energy consumption issue.

Actually, the network energy consumption as common problem exists in not only DCN but also in other ICN projects such as CCN, NetInf, DONA [15], PSIRP [16], and PUSUIT [17]. Another common problem of these ICN projects is that can only provide original data and cannot reuse kinds of service resources such as original data, processing software and services to process or generate various data to meet user’s requirements and conditions. For example, according to user device monitor size, an original image can be processed into different size or pixel on network to meet different device monitor size to display. This way not only improves service quality but also reduces the amount of images suited for kinds of devices. And meanwhile, this way also improves services utilization rate.

2.2 Ubiquitous Grid Networking Environment (uGrid)

The Ubiquitous Grid Networking Environment (uGrid) has been proposed for ubiquitous society in 2007 [18]. In uGrid, everything assigned an IP address from a device to a program is defined as Service-Part (SP). User cannot only use desired SP but also enjoy mash-up service provided by combining several SPs. In order to provide mash-up service, two core technologies Service-Routing (SR) and Service-Signaling (SS) are realized [19][20]. Service-Routing extended Open Shortest Path First Traffic Engineering extension (OSPF-TE) [21] is used for searching relative SPs and computing optimal service flow path of mash-up service. Service-Signaling extended Resource Reservation Protocol Traffic Engineering extension (RSVP-TE) [22] is used for connecting selected SPs and establishing service flow path. Then, data is transported through SPs under established service flow path. In other words, the SR and SS technologies can be applied to generate kinds of data by combining SPs on network. However, uGrid cannot support multicast communication, and doesn't take into account network energy consumption issue.

2.3 Network Virtualization Technology of JGN-X

Network virtualization technology of JGN-X aims to isolate computational resources inside network (e.g., routers and switches) act as network resources [23][24]. Its features such as topology awareness, re-configurability, resource isolation, programmability and network abstraction, make it different from traditional concepts such as Virtual Private Networks (VPNs) and overlay network. Based on network virtualization technology of JGN-X, a slice is defined as an isolated set of computational and network resources allocated and deployed across the entire network. Each slice can be considered as different network abstraction. And meanwhile these slices are isolated from each other and have no cross-talk between them. Therefore, In JGN-X, one slice as one network may be defined an arbitrary network architecture which is consisted of arbitrary protocols, data frame and message format, and rules of actions such as whether to transmit data via circuits or packets, how to route data within a slice [25][26]. It is possible that multiple network architectures and services with slices per user and even per application can be concurrently implemented on a single shared physical infrastructure such as Figure.1. Slice #1, #2, #3 are defined as different layer switched network.

Therefore, in order to solve network energy consumption issue of ICN and uGrid, reuse service resource issue of ICN, multicast communication issue of uGrid, our proposal is constructed on network virtualization platform of JGN-X. Our challenge not only applies multiple slices as network
infrastructure to construct our proposal but also research how to achieve the network energy efficiency optimization in data transmission, network topology and data distribution [27].

3. E3-DCN Architecture

We design multi-overlay networks Data Centric Overlay Network (DCON) and Data Generation Overlay Network (DGON) to realize network functions through combining them such as Figure 2. DCON is an extension of DCN and realizes a non-IP network. The uGrid concept is imported into DGON to realize data generation function. In E3-DCN, DCON handles pre-registered data and receives all users’ requests. User sends query into DCON and gets an exact matched data from DCON. If the requested data is not found in DCON, the network will automatically request DGON from DCON to generate the requested data. Finally, the generated data will be pushed into DCON and provided to user from DCON. If E3-DCN can’t provide the requested data from DCON and DGON, user will receive a “NACK” message which indicates the requested data has not been published into E3-DCN by service provider.

In order to realize the optimization of network resource and energy consumption, we adopt multi-slice networks as network infrastructure to construct our proposal based on JGN-X’s network virtualization platform. In current Internet, the overlay network service receives all data requests, and the data is transmitted via underlay network such as current peer-to-peer (P2P) overlay network application. Because current Internet’s overlay network services cannot collect detailed underlay network resources information, it cannot manage and control the underlay network. In other words, it is impossible to realize the optimization of underlay-network resources. In order to solve this problem, Generalized Multi-protocol label switching (GMPLS) user-to-network interface (UNI) signaling based P2P as an underlay network control technology has been proposed in [28]. The control technology can collect underlay network information and control the underlay network. Therefore, in order to achieve network resource optimization in our proposal, we adopt the control technology to manage and control underlay network. In addition, in JGN-X network virtualization Platform, a slice, which is composed of virtual nodes (VNs), virtual links and management system [25], has independent network and computational resources. One slice network can be used as one independent network infrastructure. And meanwhile, the slice network resource information can be collected at any time such as bandwidth, underlay physical links of virtual link, traffic situation, geographical location of VNs, and so on. Therefore, using above control technology, we can compute optimal data transmission path based on slice network according to collected network resource information. Additionally, JGN-X is an optical/electric packet/path integrated switching network [24]. Therefore, for the optimization of data transmission energy, we adopt multi-slice networks that one is packet switching slice network and another is circuit switching slice network to transmit data. And moreover, according to collected network resource information, the optimal energy network topology can be also computed and reconfigured at any time.

3.1 Multi-slice networks as Network Infrastructure

The separation of control plane from data plane is a key feature of E3-DCN architecture. The control plane collects underlay network resource information, and manages and controls underlay network such as where the traffic is sent. The data plane is used for forwarding traffic based on the result determined by control plane. Therefore, we use a slice called Control Plane Slice (CPS) as control plane network. Packet switching based on statistical multiplexing is very efficient because it can share link resource with other users. But switching energy and latency problems exist in packet switching. Path/circuit switching is low-energy and low-latency. Path/circuit switching is suitable for real-time application and large data transmission. But the circuit path established between sender and receiver must set up prior to data transmission, so it doesn’t suit for small data transmission and cannot share link resource to others. Therefore, in order to achieve the optimization of data transmission energy, we adopt two slices for data plane, one is Packet Switching Slice (PSS) and another is Circuit Switching Slice (CSS). Three slice networks as network infrastructure are shown in Figure 2. CPS is used as control plane network and exchanges kinds of routing and management message such as DCON’s routing messages,
data information, DGON’s Service-Routing and Service-Signaling protocol messages. PSS provides a main network topology of E^3-DCN. The topology of PSS should be projected to a logical topology of CPS. A Generic Routing Encapsulation (GRE) tunnel is applied between two VNs of CPS to match both topologies. CSS’s topology is not fixed and can be controlled and reconfigured dynamically. For example, an optical path is set-up or tear-down between two VNs of CSS at the beginning or end of data transmission. In E^3-DCN, data are mainly exchanged via PSS. If high bandwidth transmission is required and/or long holding time transmission is required, CSS is configured and used to bypass packet switching to reduce data transmission energy and ensure Quality of Service (QoS).

Two overlay networks DCON and DGON which are constructed on three slices. User’s query is sent to DCON, if the requested data is found in DCON, the data is transported to user via PSS or CSS. If the requested data are not found in DCON, then the query is sent to DGON. In DGON, the requested data is generated by connecting Service-Parts, and provided to DCON via PSS or CSS. Figure 3 shows a semantic diagram of E^3-DCN logical node architecture. The logical node is connected to three slices and users. Ethernet is used as a kind of access method. And Virtual Local Area Network (VLAN) technology is applied to distinguish these slices. Finally, the logical architecture will be mapped into JGN-X’s network virtualization platform.

In order to realize an energy efficient ICN, we define a brand new communication mode for DCON. In CCN, NetInf and DCN, request packets and data packets, which are equal numbers, are transmitted in hop-by-hop. This way results in a lot of useless traffic is transmitted and processing is occurred in intermediate nodes before the best data source is determined. And meanwhile, this issue can also result in energy consumption is increased especially when the user is more. In order to avoid this issue, we design a new communication model for DCON which includes four phases Exploration, Response, Request and Transmission such as Figure 5. In Exploration phase, data sources holding requested data are searched by exploration messages forwarded in hop-by-hop. When data source node receives the exploration message, data source sends response message to user based on track of exploration message. This phase is called Response. During Response, the optimal transmission path can be determined in intermediate nodes by Route Selection which is a function of DCON node. For example, the best data source’s response is the fastest, so its track as optimal transmission path is applied, and other responses are discarded at the intermediate node. In Request and Transmission phases, according to the optimal data transmission path, the request packets for data transmission are sent to optimal data source, the optimal data source sent data packets to response request.

The proposed communication model is different from TCP 3-way handshake communication model [30]. TCP 3-way handshake shakes hand with a specified host, whereas our model shakes hand with an optimal data source chosen from any number of data sources. The advantages of DCON communication model are that useless traffic, processing and energy consumption can be avoided comparing to CCN, NetInf and DCN. Moreover, when requested data cannot be directly provided from DCON, DCON will request DGON to generate the request data. Because multiple DCON nodes existed in DCON can send request to DGON for data generation, how to decide the optimal request node becomes a problem. By our communication model, this problem can be solved because the optimal request node can be determined in Response phase before requesting DGON.

3.2 Multi-overlay Networks for Network Function

- Data-centric Overlay Network (DCON)

In JGN-X, a virtual link is defined between two VNs. Therefore, each slice’s virtual link is terminated at each slice’s VN. Links between E^3-DCN node and each slice shown in Figure 4 are terminated at the VN within each slice. As a result, E^3-DCN node is composed of three VNs. Three VNs are connected via a network constructed in the real world. In JGN-X, a link between the VN and the real world is made by a network connector (NC). Figure 4 shows a composite structure diagram of E^3-DCN node including NC and three VNs. The VN of CPS has main for E^3-DCN node, other two VNs have a role in Ethernet Bridge or Ethernet switch which connects virtual network in the slice and the Ethernet switch in the real network [29].
We define four type packets to realize our communication model. They are Exploration Packet, Response Packet, Request Packet and Data Packet. They all contain data name item. Exploration Packet as exploration message is used for exploring data source. Response Packet as response message is used for responding exploration. And it contains a cost item shows the minimum transmission energy cost from data source to current intermediate node. Exploration Packet and Response Packet as routing message is routed in CPS. And they are used in Exploration and Response phase to decide the optimal data transmission path for data request and transmission. Request Packet as request message is sent to optimal data source. And Data Packet is used for carrying data block from optimal data source to user. Request Packet is routed in CPS. Data Packet is transmitted in PSS or CSS.

DCON is overlaid on CPS holds topology and link status information of PSS and CSS. Therefore, routing topology of DCON can be abstracted as adjacent nodes are connected with one or two links. The case of two links shows one is PSS’s link and another is CSS’s. The case of one link shows only PSS or CSS link exists. Each link has a weight that indicates data transmission energy cost. DCON routing includes Exploration and Response phases. We design three tables to realize DCON routing. They are Backtrack Query Table (BQT), Forwarding Query Table (FQT) and Link Cost Table (LCT) shown in Figure 6. LCT maintains link cost as weight used for computing the optimal transmission path. In Exploration phase, BQT is used for recording track where exploration message comes from. FQT records exploration message’s forwarding directions where go to. In Response phase, the minimum data transmission energy cost from data source to current intermediate node is hold by Response Packet. Based on LCT, the next hop which link is used among PSS and CSS is computed by Route Selection. Finally, the result is reflected into Response Packet sent to adjacent node based on BQT and FQT, and meanwhile updated into BQT and FQT for next Request phase.

DCON node model is shown in Figure 6. It contains four engines Extracting Engine, Analyzing Engine, Forwarding Engine and Caching Engine, four directories G-DCON, G-DGON, L-DCON and L-DGON Directory (G: Global, L: Local), three routing tables Backtrack Query Table (BQT), Forwarding Query Table (FQT) and Link Cost Table (LCT) described on above. Extracting Engine is mainly responsible for extracting information from user’s query. The extracted information such as data name is delivered to Analyzing Engine to judge whether the requested data exists and whether can be provided from current node. Then, based on the judged result, Forwarding Engine continues to explore the requested data, or responds the data exploration, or provide requested data from current node. The Route Selection included in Forwarding Engine is used for determining the optimal transmission path. Caching Engine is responsible for caching data. G-DCON keeps all data information of DCON which shows these data can be directly provided from DCON. G-DGON keeps the information of data generated in DGON. When the request data information does not exist in G-DCON and GDGON, the current node as Sorry Server sends the “NACK” message to response user [31]. L-DCON keeps local data information that shows what data can be directly provided from current node. And L-DGON keeps generated data information. If the requested data exists in L-DGON, Forwarding Engine requests DGON to generate the requested data according to data name. Least Recently Used (LRU) is applied into directories for quick retrieval [32]. Data Transceiver is responsible for receiving and forwarding packets from every interface such as V#CPS.1, V#PSS.1, V#CSS.1 shown in Figure 6. It can communicate with other engines. Figure 6 is shown an E³-DCN node holds three branches and adopts virtual interfaces to connect three slice network based on VLAN technology.

- **Data Generation Overlay Network (DGON)**

DGON is responsible for data generation based on Service-Routing and Service-Signaling technologies. Its routing topology is the same as DCON’s that there are one or two links between adjacent nodes. The optimal data transmission path connected with several Service-Parts can be computed based on the routing topology.

DGON node model is shown in Figure 6. It is composed of uGrid Engine, Caching Engine and Generation Data Table. uGrid Engine includes Service-Routing and Service-Signaling modules. Generation Data Table keeps data generation information such as data name and Service-Part information. The request of data generation is sent from DCON when DCON can’t directly provide the requested
data. DGON extracts data name based on data generation request. According to data name, Service-Part information is resolved based on Generation Data Table. Then, uGrid Engine performs Service-Routing to get the optimal data transmission path, and executes the Service-Signaling to establish connection connected selected Service-Parts to generate data. Next, the generated data is cached by Caching Engine and its information is registered into DCON. At last, the generated data is provided to user from DCON. Data Transceiver is the same as DCON’s.

### 3.3 Overview of E³-DCN Service

The overview of E³-DCN service is shown in Figure 7 (The unused links are not drawn in DCON and DGON). E³-DCN node connects with three slices and includes two overlay network nodes. Service provider publishes data “a.avi” into E³-DCN using `publishData` Application Programming Interface (API). The “a.avi” can be generated by combining Service-Parts A and B. Its data information is registered into GDT of DGON and directories of DCON. User1, User2 and User3 request the same data “a.avi” using `getData` API. Their queries are sent to DCON. The data generation request is sent to DGON from node5 of DCON. The optimal data source and transmission path of DCON is determined based on proposed communication model. And the optimal data transmission path finally forms a multicast tree to transmit “a.avi” from node5 of DCON. The energy efficiency of data transmission path is based on CSS and PSS. DGON’s data transmission path is changed at node2 from CSS to PSS, and DCON’s data transmission path is changed at node4 and node3.
4. Energy Optimized Strategies in E$^3$-DCN

There are different level approaches they are device-level, equipment-level and network-level can be applied to realize energy efficient network [33]. About E$^3$-DCN, we mainly focus on network-level approach which is used for achieving dynamic network-level energy efficiency optimization such as topology and route optimization, data/service relocation and caching optimization, reducing traffic peak and optical switching for bulk traffic. Therefore, there are three energy optimized strategies are applied to E$^3$-DCN.

- **Strategy I: Dynamic Network Reconfiguration**

Initially, a network topology is designed based on some requirements such as minimum transport delay, minimum network resources, accommodate traffic demand, and ensure resiliency. The network is usually designed to endure the maximum traffic demand, but traffic demand is dynamically changed in practice. In case of the lower traffic demand, energy efficient traffic engineering (TE) [34] [35] [36] can be applied to concentrate the traffic into limited number of links and nodes, and then unused links and nodes are shutdown to save operating power consumption. In E$^3$-DCN, the energy efficient TE is applied to virtual networks. The topology of virtual network is dynamically reconfigured to contain the minimum number of nodes and links. To realize the operation energy saving in the real network, JGN-X’s network virtualization platform should support real node and link’s shutdown operation. Service-Copy is another dynamic reconfiguration approach [37]. A Service-Part, which provides processing function i.e. software, can be easily copied from a computer to other computers. Virtual Machine (VM) copy/migration is also an example of the Service-Copy. Though Service-Copy consumes additional power consumption, there are some advantages as follows: 1) Shorter data transmission route can be formed by coping Service-Part. Data transmission and switching energy consumption can be also reduced. 2) The copied Service-Part can be shared for other users. 3) By Service-Copy, the same Service-Parts are distributed multiple places, so that network congestion can be voided in some degree. 4) It is not need to add the new computer for copied Service-Part.

- **Strategy II: Circuit Switching Bypass**

In general, data transmission cost of packet switching network is cheaper comparing to circuit switching network. This is because the bandwidth can be shared among traffic flows based on statistical multiplexing in packet switching network. However, the routing and transmission delay issue exists in packet switching. Therefore when a large data is transmitted or strict quality of service (QoS) preservation is required, circuit switching is preferred, and packet switching is avoided in as far as possible. This is called “circuit switching bypass” [38] [39]. In E$^3$-DCN, links in DCON and DGCON are provided via CSS and PSS. A link of CSS is composed of link termination Ethernet switches, transmission links, and circuit switches. A link of PSS is composed of link termination Ethernet switches, transmission links, and packet switches. In [40], it has been shown that an optical circuit switch requires 0.5 nJ/bit switching energy and electrical packet switch such as an Ethernet switch and an IP router requires 10 nJ/bit switching energy. The link of PSS is more energy consumption, but its transport cost is cheaper. On the contrary, the link of CSS requires less energy consumption but its transport cost is higher. Therefore, to determine which links should be used, an energy consumption estimation algorithm is required. And for transmitting data by optimal transmission path, an optimized energy routing algorithm is also required. The algorithm can compute energy optimized path according to some parameters such as data size, data transmission speed, flow duration time, and some QoS items.

- **Strategy III: Caching Optimization**

An excellent caching strategy can not only shorten the response time but also reduce the routing and transmission distance between data and user so as to achieve better energy saving. And meanwhile, data caching inside network can also reduce the total network load and avoid network congestion [41] [42]. Because E$^3$-DCN can be used as a CDN application, data caching strategy becomes a very important issue. We are studying the optimized data caching strategy for DCON and DGON. And we will also consider to extend Open Shortest Path First (OSPF) and Border Gateway Protocol (BGP) for sharing cached data among autonomous systems (AS)’s cache router to achieve scalability.

5. Evaluations

The energy consumption of transmission and network performance of DCON were evaluated. We simulated three networks DCON, CCN/DCN/NetInf (CDNI) and TCP/IP. The parameters of network topology used in simulations are shown in Table I.

<table>
<thead>
<tr>
<th>Table I. The Parameters of Simulations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Topology Nodes</strong></td>
</tr>
<tr>
<td><strong>Topology Edges</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Cost of CSS’s link</strong></td>
</tr>
<tr>
<td><strong>Cost of PSS’s link</strong></td>
</tr>
<tr>
<td><strong>Request Users</strong></td>
</tr>
<tr>
<td><strong>Data Source</strong></td>
</tr>
</tbody>
</table>

Every network topology contains 5000 nodes. DCON takes into account the optimization of energy consumption for data transmission based on PSS and CSS, so there are two edges one is PSS link and another is CSS link exist in between adjacent nodes of DCON. Because CCN/DCN/NetInf and TCP/IP only uses PSS or CSS to transmit data and doesn’t take into account data
transmission energy computation problem, there is only one PSS or CSS link exists between adjacent nodes of its topology. Therefore, the amount of edges of network topology used in CCN/DCN/NetInf and TCP/IP is 34972, used in DCON’s is 69944. The link cost that indicates data transmission energy consumption is random positive integer. Its value range is between 1 and 100. Only one data source exists in each network. The amount of request user increases from 100 to 2500.

We evaluated data transmission energy consumption cost by comparing DCON with CCN/DCN/NetInf networks. Because CCN, DCN and NetInf all use name-based routing, they are considered as the same type network in simulations. The network performance is evaluated by comparing with TCP/IP network. Because TCP/IP network supports point-to-point communication channel, with the increasing of user, the used links of channels between adjacent nodes will be increased. Because DCON realizes ICN and can aggregate the same requests to one in intermediate node, there is only one link used between adjacent nodes when all users request for the same data. Therefore, the network performance is evaluated based on the amount of used links. The number of used links is more that indicates the network performance is lower, because network congestion and delay problem occur easily.

The evaluation of data transmission energy consumption is shown in Figure.8. The energy consumption cost of DCON is less than CCN/DCN/NetInf about 40%. The result shows that DCON uses the optimal energy route to transmit data in contrast to CCN, DCN and NetInf. This is because DCON can choose the minimum energy consumption link for transmission based on PSS and CSS. CCN, DCN and NetInf don’t take into account energy consumption problem in data transmission and only compute data transmission route based on PSS or CSS network. Therefore, the total data transmission energy consumption of DCON is less than of CCN/DCN/NetInf.

6. Experiments

In order to validate our proposal, we have realized non-IP communication network DCON, data generation of DGON and data transmission using PSS and CSS.

The transmission mechanism of DCON is shown in Figure.10. Based on raw socket technology that can directly send and receive data from data link layer with specified interface [43], we defined data packet encapsulated in Ethernet frame including items Data ID, Data Block. Because VLAN technology is applied to distinguish PSS and CSS slices, so VLAN Tag is also encapsulated in frame. Based on this mechanism, we demonstrated DCON network composed of 6 DCON nodes with Ethernet and optical switches and exhibited in the 5th Symposium of New Generation Network in Tokyo of Japan [44], MPSL2012 international conference [45] and KEIO TECHNO-MALL 2012 [46].
Service-Routing and Service-Signaling technology were imported into DGON for data generation through combining Service-Parts. In Cloud Symposium of Japan 2013 [47], E3-DCN prototype, shown in Figure.1, was composed of DCON and DGON overlay networks based on JGN-X. DGON realized data generation by combining Sendai data center in Tohoku area of Japan, Keio University in Kanagawa Japan, and Cloud Symposium Place in Tokyo Japan, three places’ Service-Parts. DCON only deployed in Cloud Symposium place because it is non-IP network. The generated data of DGON was pushed into DCON and then transmitted to users.

7. Conclusions and Future Work

We have proposed the architecture of E3-DCN based on JGN-X network virtualization platform in this paper. E3-DCN is composed of two overlay networks DCON and DGON, and three slice networks CPS, PSS and CSS. DCON realizes an ICN. DGON, which is data generation network, not only can generate requested data for DCON by combining Service-Parts, but also can improve resource utilization rate. CPS as control plane is used for control network. PSS and CSS are used for data transmission. E3-DCN not only solves problems existed in CCN, DCN, NetInf and uGrid but also achieves an energy efficient ICN network by reconfigure network topology, circuit switching bypass and data caching. By evaluations, the optimization of data transmission energy efficiency can be achieved based on PSS and CSS comparing to other ICN projects. And the network performance is better than current TCP/IP network. Furthermore, we have developed the prototype of E3-DCN to validate our proposal on JGN-X.

About future works, we will continue to study optimal energy algorithms for data transmission, dynamic network reconfiguration, circuit switching bypass and optimal data caching. And we will also study and develop E3-DCN APIs to get underlay network information for energy algorithms and network management. Finally, we will integrate these energy optimized algorithms and network management with prototype of E3-DCN on JGN-X.

Acknowledgments

This research is supported by “Energy Efficient and Enhanced-type Data-centric Network” Project funded by the National Institute of Information and Communication Technology (NICT).

References

192, OECD Publishing.


Shanming Zhang received the B.S. degree in computer science from Changchun University of Technology, China, in 2006. He joined Sony Global Information System (Dalian) Co., Ltd, China, in 2006. He was engaged in the development of information system and maintenance. He enrolled and received the M.E. degree in electronics engineering from Keio University in 2011 and 2013, respectively. He is currently working toward the Ph.D. degree in Graduate School of Science and Technology, Keio University, Japan. His research interests include Information-centric networking, network functions virtualization, network energy efficiency and data caching. He is a member of the IEICE and a student member of the IEEE.

Hidetoshi Takeshita graduated from Shizuoka University, Japan where he received the B.E in 1974. In 1974 he joined NEC Corporation, Tokyo, Japan, where he engaged in the development of electronic switching systems, digital switching systems, and system engineering for fixed and mobile carriers. He graduated from Keio University where he received the Ph.D. degree in Sept. 2012. From 2010 to 2012, he was a Research Assistant of Keio University Global COE Program, and from June to September in 2012, he was a researcher in Graduate School of Science and Technology of Keio University. Since Oct. 2012, he has been a project assistant professor. He is a member of the IEEE, and the IEICE.

Satoru Okamoto is a Project Professor at Keio University, Yokohama, Japan. He received his B.E., M.E. and Ph.D. degrees in electronics engineering from Hokkaido University, Hokkaido, Japan, in 1986, 1988 and 1994. In 1988, he joined Nippon Telegraph and Telephone Corporation (NTT), Japan, where, he conducted research on ATM cross-connect system architectures, photonic switching systems, optical path network architectures, and participated in development of GMPLS-controlled HIKARI router ("photonic MPLS router") systems. He joined Keio University in 2006. He is now researching future IP + optical network technologies, and application over photonic network technologies. He was the chair of the IEICE Technical Committee on Photonic Network (PN) (2010-2011). He is an IEICE Fellow and an IEEE Senior Member.

Naoaki Yamanaka graduated from Keio University, Japan where he received B.E., M.E., and Ph. D. degrees in engineering in 1981, 1983 and 1991, respectively. In 1983 he joined Nippon Telegraph and Telephone Corporation’s (NTT’s) Communication Switching Laboratories, Tokyo, Japan, where he was engaged in the research and development of a high-speed switching system and technologies for Broadband ISDN services. Since 1994, he has been active in the development of ATM-based backbone networks and system. He is now researching the future optical IP network, and optical MPLS router systems. He is currently a professor of Keio University. He is a vice director of Asia Pacific Board at IEEE Communications Society. He is an IEEE Fellow and an IEICE Fellow.
Information for Authors

Scope

The aim of the International Journal of Computer & Information Science is to present information on well defined theoretical results and empirical studies, to describe current and prospective practices and techniques, to provide critical analysis of the use of CASE tools and to present broad issues in the fields of Computer Science and Engineering, Software Engineering and Information Science and Engineering. A primary aim of the journal will be to make the research accessible and useful to practitioners through carefully written methodology section and discussions of research results in ways that relates to practitioners.

Submission Policy

1. The IJCIS does not accept manuscripts or works that have been published in a journal, but accepts substantially revised and expanded versions of papers that were previously presented at conferences. In this case, the author should inform the Editor-in-Chief of the previous publication when the paper is submitted.

2. Submissions, referencing and all the correspondence will be conducted by e-mail.

3. Manuscripts should be in English and are subject to peer review before selection.

Procedure for Submission for Review

1. Send your complete text of manuscript to the Editor-in-Chief (e-mail: lee@cps.cmich.edu) via e-mail as an attachment document either in Word, PDF or PostScript format.

2. It may be useful for non-English speakers to have a native English speaker read through the article before submission.

3. Your e-mail message (to the Editor-in-Chief) should include the title of your paper, author(s) information (department/section, university/company, phone/fax numbers, e-mail address), the corresponding author’s complete address (mailing address), and area(s) of the paper.

Style of Manuscript to be Submitted for Review

1. The text should be double-spaced, two columned, single-sided in Times 12-point type on 8.5-inch x 11-inch paper with margins of 1 inch on all four sides. It should not exceed 30 double-spaced pages. All pages must be numbered sequentially.

2. Provide a 200-250 word abstract, 5-7 keywords. Your abstract should briefly summarize the essential contents.

3. All figures/tables must be captioned, numbered and referenced in the text.

4. An acknowledgement section may be presented after the conclusion section.

5. References should appear at the end of the paper with items referred to by numbers in square brackets.

Journal Papers: author’s first initials, followed by last name, title in quotation marks, journal (in Italics), volume and issue number, inclusive pages, month and year. For e.g.:


Papers of Proceedings: author’s first initials, followed by last name, title in quotation marks, proceedings / symposium / workshop name (in Italics), location, month, and year, and inclusive page numbers. For e.g.:


Books: author’s first initials, followed by last name, title (in Italics), edition, location: publisher, year. For e.g.:


Accepted Manuscript Submission

All accepted papers are edited electronically. You are required to supply an electronic version of the accepted paper. The detailed guidelines for preparing your final manuscript will be sent upon article acceptance. It is also available on the ACIS website at http://acisinternational.org.

Page Charges

After a manuscript has been accepted for publication, the author will be invoiced a page charge of $400.00 USD for up to 10 pages if he/she is not a member of ACIS and thereafter an additional $30 per printed page to cover part of publication after the receipt of your final manuscript (camera ready).