International Journal of Innovative
Computing, Information and Control
Volume 8, Number 5(A), May 2012

ICIC International ©2012 ISSN 1349-4198
pp. 3115–3134

A CONTEXT ANALYSIS SCHEME OF DETECTING PERSONAL AND CONFIDENTIAL INFORMATION

Toshihiro Satomi[1], Atlam El-Sayed[1,2], Kazuhiro Morita[1]
Masao Fuketa[1] and Jun-ichi Aoe[1]

[1]Department of Information Science and Intelligent Systems
University of Tokushima
2-1 Minamijyousanjima-cho, Tokushima 770-8506, Japan
atlam@is.tokushima-u.ac.jp

[2]Department of Statistics and Computer Science
Faculty of Science
Tanta University
El-Giesh St., Tanta, Gharbia, Egypt
satlam@yahoo.com

Received March 2011; revised August 2011

Abstract. *The frequent information exchange has been spread widely in the Internet making communication convenience to a huge number of people. However, there are many problems related to stolen personal and confidential (PC) information through the Internet and resources in organizations. In order to support the supervision of PC information, there are many PC texts depending on detection systems. The traditional methods depend on words or a sequence of words without considering the context analysis, and many irrelevant candidates of possible PC information are extracted. This paper presents a new detection scheme for non-PC texts by using context analysis. The heart of the new approach is introducing neglect (NEG) expressions that can cancel the detected PC information. It enables us to reduce extra-detections or human efforts for non-PC texts. Experimental results for context data show that the improvement of the presented method becomes 66.5% compared with the traditional methods. Moreover, the human efforts reduce by about 80% comparing to by using the traditional methods.*
Keywords: Personal and confidential information, Multi-attribute matching, Neglected expressions

1. **Introduction.** Recently, the frequent information exchange has been spread widely in the Internet. This information is used as basic services through the website to exchange E-mails, to buy and sell products on websites, to keep personal information on blogs, and to exchange ideas and opinions on the social network service (SNS). However, there are many problems related to stolen personal and confidential (PC) information such as name, address, medical examination, credit card, bank account, E-mail, patent.

For example, in 2005, information on more than 40 million credit cards was stolen from MasterCard International [1]. In 2007, ATM PINs were stolen from Citibank ATMs and the alleged thieves stole about 2 million dollars by transmitting from the ATMs to the transaction processing computers [2]. In 2009, Yahoo, Gmail and Hotmail accounts were breached by a phishing scam, therefore the hacked web mail accounts used for sending spam [3].

Based on these problems, many countries have been establishing new laws related to personal and confidential information as in the following three categories:

a) **For Private information:** the Data Protection Directive (DPD) [4] was established in a European Union directive, and the Personal Information Protection Law (PIPL) [5] was established to protect individuals' rights in Japan.
b) **For Medical information:** the Children's Online Privacy Protection Act (COPPA), and the Health Insurance Portability and Accountability Act (HIPAA) were established in USA [6].
c) **For Financial information:** Protection and Electronic Documents Act (PIPEDA) was established in Canada [6] and the Fair Accurate Credit Transactions Act (FACTA) was established in USA [6].

Although statistics-based filtering [7-9] is the popular technologies, it is difficult to detect the precise locations for expressions. Other researches [10-13] introduced content analysis for harmful websites such as pornography, drug, violence, crime, but the scheme is restricted by using one-class SVM classification [14] in the context analysis to detect harmful words. Kadoya [15] proposed an e-mail filtering scheme depending on rule based knowledge approach, but there is no discussion in Kadoya scheme about context analysis. There are many extra-detection results from non-PC texts, because traditional schemes are based on sentences analysis or non-context analysis. Therefore, it is very important to reduce the extra-detection or human efforts by proposed context expressions analysis for PC detection.

In order to solve the current problems, this paper presents a new context filtering algorithm to reduce human efforts and to improve the accuracy rate for non-PC texts. The detection method is based on rule-based knowledge and it defines separate co-occurrence (SC) expressions that cannot be detected by word sequences of the traditional methods. The context analyses for SC expressions can be performed by a two-phase process using multi-attribute rules. The heart of the new approach is to introduce neglect (NEG) expressions that can cancel the detected PC information. It enables us to reduce extra-detection or human efforts from non-PC texts. By the experimental results for context data, it turns out that the improvement of the presented context analysis method becomes 66.5% to the traditional methods. Moreover, the human effort reduces by about 80% than by using the traditional method.

Section 2 introduces PC and NEG concepts. Section 3 proposes how to define multi-attribute rules for detecting PC information together with examples. Section 4 presents a context analysis algorithm by two-phase multi-attribute matching that can utilize NEG concepts. Section 5 evaluates the presented method by experimental results. Section 6 describes conclusion and possible future work.

2. Personal and Confidential Expressions.

2.1. Outline of the presented method.

Consider the examples shown in Figures 1 and 2 to explain the concept of the presented method. Figure 1 represents text A including PC information, however, Figure 2 represents Text B including no PC information, where <> means hierarchical concepts for expressions and the detail will be explained in the whole paper.

In *(a1)*, *(a2)* and *(a6)* of Text A in Figure 1 includes name "*Sara White*", birthday "*November 14th, 1965*" and serious medical information as "*liver inflammation*", "*40 IU in GPT*", "*AIDS*". Therefore, we can say that Text A has PC information. On the other hand, although Text B in Figure 2 includes the same expressions of Text A, it is clear that Text B has no PC information.

(a1) Name: Sara White <NAME\HUMAN>
(a2) Birthday: November 14th, 1965<DATE\BIRTHDAY>
(a3) Occupation: Director <TITLE\JOB> of Johnson company <NAME\COMPANY>
(a4) Examination <EXAMINATION>
(a5) Blood type is Rh^- <MEDICAL\BLOOD TYPE>
(a6) She <PERSON\DEIXIS> might be acute liver inflammation <MEDICAL\DISEASE>
(a7) due to over 40 IU in GPT <MEDICAL\TEST\RESULTS>
(a8) and includes serious sickness by AIDS <MEDICAL\DSEASE>

FIGURE 1. Text A including PC information

(b1) Dr. Sara White<NAME\HUMAN>
(b2) November 14th, 2009<DATE\BIRTHDAY>
(b3) Director < TITLE\JOB> of Johnson Hospital<NAME\COMPANY>
(b4) Webpage News:
(b5) Attention about blood type Rh^- <MEDICAL\BLOOD TYPE> can't receive blood from Rh^+<MEDICAL\BLOOD TYPE> donor.
(b6) Visit to hospital home page <INTRODUCTION>.
(b7) if you want to see <REQUEST>acute liver inflammation that <MEDICAL\DISEASE> it is due to over 40 IU in GPT <MEDICAL\TEST\RESULTS>
(b8) and if you hope to get information about AIDS <MEDICAL\DISEASE>

FIGURE 2. Text B including no PC information

In the presented method, a two-phase process is introduced to solve the differences between Texts A and B. The first phase is to determine essential elements of PC expressions and the second phase is to decide the final results whether the input text is PC or not.

Consider the following sentences in Text A. In *(a1)*, *"Name: Sara White"* means *"human name"*, denoted by concept <NAME\HUMAN> which is the sub-concept of super concept <NAME> including building names, product names, and so on. In *(a8)*, *"AIDS"* is denoted by concept *<MEDICAL\DISEASE>*. These concepts are detected by the first phase as the basic elements of PC expressions and the second phase determines the final results by using (SC) rules such as <NAME\HUMAN> + *<MEDICAL\DISEASE>*, where '+' represents co-occurrence relationships among concepts and expressions. That is to say, the context analysis is carried out by this second phase.

For Text B, the presented method defines neglect (NEG) concepts that can delete PC information detected in the first phase, and this decision is also performed by the second phase as context analysis. Consider *(b7)* of Text B in Figure 2. In *(b7)*, "*if you want to see*" means that of "request", denoted by *<WEB PAGE>*. Therefore, the PC expression "*acute liver inflammation*" *<MEDICAL\DISEASE>* can be neglected by (SC) rule *<REQUEST>* + *<MEDICAL\DISEASE>*.

In this paper, the presented method (PM) is a new detection scheme based on context analysis to improve the accuracy rate by reducing the extra-detections for non-PC texts. In order to estimate the efficiency of the presented method, the traditional method (TM) is introduced as a scheme based on sentences analysis or non-context analysis.

2.2. **Concepts for personal and confidential information.** PC information means private secret information that must be protected. Concepts of PC information are basic

elements that extract a variety of expressions and the rule-based knowledge to be presented utilizes the concepts.

(1) Concept <NAME\HUMAN>

The core of PC information is the concept <NAME\HUMAN> for a person name, but it is generally ambiguous in expressions of sentences because name is utilized for some kinds of names such as company, street [16], building names and so on. Therefore, it is important to solve this ambiguity by using concept <TITLE> as follows:

(a) **<TITLE\HONORIFIC>** means Mr., Ms. and Miss for conferring or showing respect or honor. Consider the following example sentence.
"Ms. <TITLE\HONORIFIC> Johnson <NAME> was very angry last night".
"*Johnson*" is just the concept <NAME>, but it can be determined as <NAME\HUMAN> by rule <NAME\HUMAN>=<TITLE\HONORIFIC>+<NAME>.

(b) **<TITLE\JOB>** is a list of the general tasks, or functions, and responsibilities of a position. Typically, it also includes to whom the position reports, specifications such as the qualifications needed by the person in the job, salary range for the position, and so on. We have three sub-categories of job titles that are represented as follows:

(b1) **<TITLE\JOB\EXECUTIVE>** includes Chairman, Vice chairman, President, Representative Director, Director, Executive Director and Managing Director. An example sentence is shown as follows:

"Cathy<NAME> is director <TITLE\JOB\EXECUTIVE> at the ABC company".
<NAME\HUMAN> = <NAME> + <TITLE\JOB\EXECUTIVE>.

(b2) **<TITLE\JOB\OCCUPATION\PROFFESIONAL>** includes Accountant, Computer Programmer, Dentist, Engineer, Farmer, Nurse, Salesman, Web Designer and Doctor, where the notation of <OCCUPATION\PROFFESIONAL> means <PROFFESIONAL> is the subcategory of <OCCUPATION>. An example sentence is shown as follows:
"Davis <NAME> got a job as an engineer <TITLE\JOB\OCCUPATION\PROFFESIONAL> at the EFG chemical".
<NAME\HUMAN> = <NAME> + <TITLE\JOB\OCCUPATION\PROFFESIONAL>.

(b3) **<TITLE\JOB\OCCUPATION\ACADEMIC>** includes Professor, Assistant Professor, Lecturer and Teacher. An example sentence is shown as follows:
"Sara<NAME> who is a professor <TITLE\JOB\OCCUPATION\AC ADE MIC> in her department",
<NAME\HUMAN> = <NAME> + <TITLE\JOB\OCCUPATION\ACADEMIC>.
Determining <NAME\HUMAN> needs to two more SC concepts introduced in this sub-section and is a very important task for PC information.

(2) **<ADDRESS>** The concept <ADDRESS> has two kinds as follows:

(a) **<ADDRESS\LIVING>** means the place where a person or organization can be found or communicated with. <ADDRESS> includes Country (America, France, Japan, Canada) [17], City (Los Angeles, Orland, Vancouver, Calgary), Avenue, street, road, drive (Miller Road, Main Street, 82nd Avenue, Harry Drive), State, Province, Prefecture (MI (Michigan), CA (California), AB (Alberta), BC (British Columbia), Tokyo, Tokushima), Building (ABC Building,

ABC Bldg), home number (2130, 212, 65, 4), and Zip Code (10466 (U.S.A.), V5K 1E9 (Canada), 7710006 (Japan)). An example sentence is shown as follows:
"John<NAME> lives in room 201 of Marry<NAME> Building",
In this example, *"John"* and "*Marry*" is general concept <NAME>, but their ambiguities are solved by
<NAME\HUMAN> = <NAME>+"*lives*",
<ADDRESS\LIVING> = "*room 201*" + <NAME>, and
<ADDRESS\LIVING> = <NAME> +"*Building*",
where "*live*", "*room 201*", and "*Building*" are defined by concepts in the practical rule base knowledge if they have the corresponding concepts.

(b) <**ADDRESS\E-MAIL**> identifies a location to which messages can be delivered. An example sentence is shown as follows:
"Prof. John asked his students to send their reports by mail to john@example.com <ADDRESS\E-MAIL>".
It is easy to determine E-mail address in general.

(3) <**NUMBER**>

(a) <**NUMBER\PHONE**> means the number for calling a particular telephone [18]. <Phone Number> includes Country code (+1 (America), +20 (Egypt), +44 (United Kingdom), +81 (Japan)), Area code (202 (Washington D.C.), 416 (Toronto), 03 (Tokyo)), and individual phone number (2345-3456, 34-2345, 456-789). An example sentence is shown as follows:
"Please call this number, 234-4567<NUMBER>".
<PHONE NUMBER> = "call"+ <NUMBER>.

(4) Concept <DATE\BIRTHDAY>
<DATE\BIRTHDAY> means an anniversary of the day on which a person was born or its celebration. An example is shown as follows:
"Sara's<NAME> birthday is on November 14th, 1965<DATE>",
<NAME\HUMAN> = <NAME> + "*birthday*"
<DATE\BIRTHDAY> = "*birthday*"+ <DATE>.

(5) <**MEDICAL**>
Medical information classifies into three sub-categories <MEDICAL\BLOOD TYPE> [19], <MEDICAL\CONDITION> and <MEDICAL\DISEASE> [20] that are represented as follows:

(a) <**MEDICAL\BLOOD TYPE**> means any of the four main types into which human blood is divided as A, B, AB and O. Blood types are based on the presence or absence of specific antigens on red blood cells and human blood is divided in details; negative Rh^- and positive Rh^+. An example is shown as follows:
"John's blood type <BLOOD TYPE> is A<TYPE>",
<MEDICAL\BLOOD TYPE> = < BLOOD TYPE> + <TYPE>.

(b) <**MEDICAL\CONDITION**> means an assessment of the animal's weight by age and weight for height ratios, and its relative proportions of body function <Body condition> includes blood pressure (high blood pressure, 140/90, low blood pressure, and 92/57), liver function (GOT and GPT), cholesterol (CHOL) and so on. An example is shown as follows:
"She<PERSON\DEIXIS> might be acute liver inflammation <MEDICAL\ DISEASE>", <MEDICAL\CONDITION> = <PERSON\DEIXIS>
+ <MEDICAL\DISEASE>.

(b1) <**MEDICAL\CONDITION\GTP**>

"40 IU<VALUE> in GPT<MEDICAL\CONDITION\GTP>" is defined by <MEDICAL\CONDITION> = <VALUE> + <MEDICAL\ CONDITION\GTP>.

(b2) <**MEDICAL\CONDITION\GOT**>
"60 IU<VALUE> in GOT<MEDICAL\CONDITION\GOT>" is defined by the same manner of the above b1).

(b3) <**MEDICAL\DISEASE**> means a pathological condition of a part, organ, or system of an organism resulting from various causes such as infectious disease (AIDS, Malaria, and Tuberculosis). An example sentence is shown as follows:
"Sara<NAME> includes her serious sickness by AIDS<MEDICAL\DIS EASE>"
<MEDICAL\DISEASE> = <MEDICAL\DISEASE>
<MEDICAL\DISEASE> = <NAME> + <MEDICAL\DISEASE>
"She <PERSON\DEIXIS> might be acute liver inflammation <MEDICAL\DISEASE>"
<MEDICAL\DISEASE> = *<PERSON\DEIXIS>* + <MEDICAL\ DISEASE>

(6) <**ACADEMIC**>

(a) <**ACADEMIC\RECORD**> means academic information kept on files by schools. This record includesgrades [21], test scores, and related academic materials. As for grades, there are many ways to express grades such as Grade (A+, A−, B+, B, B−, C+, C, C−, D, F), Description (Exceptional, Excellent, Very Good, Good, Satisfactory, Adequate, Marginal, Failure), and Level (Level 4, Level 3, Level, 2, Level 1). An example sentence is shown as follows:
"Rachel<NAME> got A + <GRADE> for Physics <SUBJECT> last year", <ACADEMIC\RECORD> = <GRADE> + <SUBJECT>.

(7) <**FINANCIAL** >

Financial matter includes <CREDIT CARD>, <BANK ACCOUNT> [22,23] and <STOCK> [24] as follows:

(a) <**FINANCIAL\CREDIT CARD**> means the numbers on credit cards that are not randomly assigned, each digit in the number sequence has a meaning. <Credit Card> includes card type (VISA, MASTER, AMERICAN EXPRESS), expire date (06/14), name (TARO KOJIMA), card number (1111-2222-3333-4444) and card security code (222). An example sentence is as follows:
"John's card is VISA<CARD TYPE> with number 1459-1458-8962-4578<NUM BER>, <FINANCIAL\CREDIT CARD> = <CARD TYPE> + <NUMBER>.

In the above sentence, *"1459-1458-8962-4578<NUMBER>"* is ambiguous because some digits may be *Telephone number, ID number, credit number or bank account, as so on.*

Figure 3 shows an example of Actual text from the home page http://www.forbestravelguide.com/book-trip.htm for Visa Card PC information. The underline words are represented all words related to the <FINANCIAL> category, while the bold portion represents the NG expression.

From Figure 3, we can find much PC information for Visa Card as "credit card", "finance charges" which are the sub-concept of super concept <FINANCIAL>. These concepts are detected by the first phase as the basic elements of PC expressions and the second phase determines the final results by using (SC) rules such as <FINANCIAL\CREDIT CARD> + <FINANCIAL\CAR RENTAL>, where '+' represents co-occurrence relationships among concepts and expressions. That is to say, the context analysis is carried

It's true that credit cards have become important sources of identification -- **if you want to rent a car**, for example, you really need a major credit card. And used wisely, a credit card can provide convenience and allow you to make purchases with nearly a month to pay for them before finance charges kick in.

FIGURE 3. Example of PC for visa card

out by this second phase. At the same time, we can find some neglect (NEG) expressions as in concepts "**if you want to rent a car**" that mean <INTENSION\REQUEST>.

(b) <**FINANCIAL\BANK ACCOUNT**> is a financial account with a banking institution [30], recording thefinancial transactions between the customer and the bank and the resultingfinancial position of the customer with the bank <Bank Account> includes bank name (ABC bank), branch name (New York), bank code (001), branch code (009), swift code mainly used in Japan and there are from 8 to 11 digits; AAAAJPJTXXX (AAAA means bank code, JP means country code, JT means location code and XXX means branch code), ABA No. (U.S.A., 212 545 667), IBAN has 34 digits at maximum. (e.g., GBkk BBBB SSSS SSCC; GB means countries, kk means check digits, B means bank codes, S means sort code, and C means bank account number), bank account number (1234567), and name (Jiro Mori). An example is shown as follows:
"*White*<*NAME*> *saves all her money to ABC bank*<*BANK*>, *123 456 789* <*NUMBER*>",
<FINANCIAL\BANK ACCOUNT> = <BANK> + <NUMBER>.

(c) <**FINANCIAL\STOCK**> means the capital raised by a corporation through the issue of shares entitling holders to an ownership interest (equity). <STOCK> includes shares (10 shares) ticker symbol (MSFT (Microsoft), SNE (Sony), GE (General Electric)), which is a short abbreviation for traded shares and company name (ABC Company). An example is as follows:
"Sara buys 20 shares<*SHARES*> *of the ABC Company",*
<FINANCIAL\STOCK> = "buy" + <SHARES>.
There are many concepts and relationships for *"CONTRACT DOCUMENT"* and *"PATENT DOCUMENT"* such as <CONTRACT>, <AGREEMENT>, <BUYER>, <SELLER>, <GOODS>, <CLAIM>, <SECURITY INTEREST> and <SIGNATURE>. Detail explanations are not included because the readers can build expressions taking the above definitions as starting point.

2.3. **Neglect concepts.** NEG concepts are based on intension [25] and emotion (sensibility) [26] expressions. This paper shows some examples:

<INTENSION> includes expressions with user's intensions and this paper uses the following categories:

(a) <**INTENSION\INTRODUCTION**> means a personal letter of presenting one person to another as follows:
"Please visit this shop", "A seminar will be performed tomorrow", "This software is very good for your computer".

(b) <**INTENSION\REQUEST**> means to express a desire for; ask for. Often used with an infinitive or clause as follows:
"You want to know the place", "Please submit this text", "Please come to the station at 10:00".

(c) <**INTENSION\QUESTION**> means an expression of inquiry that invites or calls for a reply as follows:
"Where is the meeting place?", "When is this seminar?".

Figure 4 shows an example of Actual text from the home page *http://www.nih.gov/research matters/july2011/07252011family.htm* or Medical PC information. The underline words are represented all words related to the <MEDICAL> category, while the bold portion represents the NEG question expression.

"We hope this new information will help educate physicians to more frequently **ask patients** these **important questions**," says lead researcher Dr. Sharon Plon of Baylor College of Medicine. "Our results are relevant for all patients, since anyone may have a change that would affect their cancer screening recommendations."

FIGURE 4. Example of NEG question information

From Figure 4, we can find much PC information for medical as "educate physicians", "College of Medicine" and "cancer screening" which are the sub-concept of super concept <MEDICAL>. These concepts are detected by the first phase as the basic elements of PC expressions. At the same time, we can find some neglect (NEG) expressions as in concepts "**ask patients**" and "**important questions**" that mean <INTENSION\QUESTION>.

(d) <**INTENSION\EMOTION**> means the part of the consciousness that involves feeling; sensibility as follows:
"This bed feels odd", "I'm sad now", "The program I got yesterday is wrong".

(e) <**INTENSION\REPLY**> means to give an answer in speech or writing as follows:
"It is OK", "I will be absent from the meeting", "I refuse this work".

(f) <**INTENSION\INVITATION**> means a spoken or written request for someone's presence or participation as follows:
"Would you like to go with me?", "Let's play together".

(g) <**INTENSION\ENCOURAGEMENT**> means the expression of approval and support as follows:
"I was sorry to hear you failed the exam", "There will be more chances", "Please do your best!".

No matter you are Doctor or not, no matter what you do, you absolutely have the **power to change**. Just **trust yourself**, then you know how to recover yourself from ADIS.

FIGURE 5. Example of NEG encouragement information

Figure 5 shows an example of Actual text from the home page *http://mobiles.maxabout.com/sms/encouragement_sms.aspx* for Medical PC information. The underline words are represented all words related to the <MEDICAL> category, while the bold portion represents the NEG encouragement expression.

From Figure 5, we can find much PC information for medical as "Doctor", "Recover yourself" and "ADIS" which are the sub-concept of super concept <MEDICAL>. These concepts are detected by the first phase as the basic elements of PC expressions. At

the same time, we can find some neglect (NEG) expressions as in concepts "**power to change**" and "**trust yourself**" that mean <INTENSION\ ENCOURAGEMENT>.

(h) <**INTENSION\ASSUMPTION**> means the act of taking possession or asserting a claim as follows:
"If it rains, the game will be called off", "If you have some questions, please ask ...".

3. Multi-attribute Detection Scheme.

3.1. **Multi-attribute rules (MULTI).** For detection of the expected expressions in natural language processing, it is important to utilize an efficient matching algorithm that can treat multi-attribute formation (morphological, syntactic and semantic).

Let A be the attribute name and let V be the value of A. Then, let R be a finite set of pairs (A, V), then R is a rule structure with the following attributes:

STR: string, or, word spelling,

CAT: category, or a part of speech,

SEM: semantic information to be defined in this paper.

Personal and confidential expressions are described by combining a variety of words, phrases, categories and semantics as follows.

The formal definition depends on the description by Kiyoi [25], while rules correspond to personal and confidential expressions. For example, multi-attribute structures of the sentence "*She might have acute liver inflammation*" are defined as follows:
N_1={(STR, "*She*"), (CAT, PRONOUN), (SEM, <*PERSON\DEIXIS*>)},
N_2={(STR, "*might have*"), (CAT, VERB)},
N_3={(STR, "*acute liver inflammation*"), (CAT, NOUN), (SEM, <MEDICAL\ DISEASE>)}.

The pre-processor can detect structure candidates which define and produce a sequence of structure candidates. Therefore, the rule matching for the above expressions can be defined by multi-attribute matching Rule(p) as follows:

$$\text{RULE } (pm < n_p) = \mathrm{R}_{p1}\mathrm{R}_{p2} \ldots \mathrm{R}_{pm}\,(m < n_p)$$

Context analysis of MULTI determines candidates of personal expressions in the text and produces results (CON, x), where CON and x represents features for context analysis. The detailed method will be discussed in the next section.

Tables 1 and 2 show multi-attribute rules for the first phase of texts A and B in Figures 1 and 2, respectively, where (STR, *"name"*), (STR, *"birthday"*) and (STR, *"blood type"*) are used as examples of the attribute STR although it is defined in the practical rules, and (CAT, NOUN) for *"Sara White"* is also used as examples of the attribute CAT. Note that expression *(b7)* in Table 2 needs two rules.

Sentence *(a6)* in Table 1, the following Rule (6) detects "*She might have acute liver inflammation*".

Rule (6) = R_{61} R_{62} R_{63}
R_{61} = {(SEM, <PERSON\DEIXIS>)}
R_{62} = {(CAT, VERB)}
R_{63} = {(SEM, <MEDICAL\DISEASE>)}

Let N$\supseteq$R be the operation that N includes R. Matching is success if the following conditions are satisfied.

$\mathrm{N}_1 \supseteq \mathrm{R}_{61}$, $\mathrm{N}_2 \supseteq \mathrm{R}_{62}$, $\mathrm{N}_3 \supseteq \mathrm{R}_{63}$

The formal matching algorithm will be explained in the next subsection.

Tables 3 and 4 show the parts of multi-attribute rules on the second phase for texts A and B in Figures 1 and 2, respectively.

TABLE 1. Rules of the first phase for text A in Figure 1

Expressions	*Rules*	*Output of the first phase*
(a1) Name: Sara White	{(STR, *"name"*)} {(CAT, NOUN)}	{(CON, <NAME\HUMAN>)}
(a2) Birthday: November 14th, 1965	{(STR, *"birthday"*)} {(SEM, <DATE>)}	{(CON, <DATE\BIRTHDAY>)}
(a3) Occupation: Director of Johnson company	{(SEM, <JOB\TITLE>)} {(SEM, <NAME\COMPANY>)}	{(CON, <JOB\TITLE>)}
(a4) Examination	{(SEM, <EXAMINATION>)}	{(CON, < EXAMINATION >)}
(a5) Blood type is Rh-.	{(STR, *"blood type"*)} {(SEM, <BLODTYPE VALUE>)}	{(CON, <MEDICAL\BLOODT YPE>)}
(a6) She might have acute liver inflammation	{(SEM, <PERSON\DEIXIS>)} {(SEM, <MEDICAL\DISEASE>)}	{(CON, < MEDICAL\DISEASE>)}
(a7) due to over 40 IU in GPT	{(SEM, <VALUE>)} {(SEM, <MEDICAL\TEST>)}	{(CON, <MEDICAL\TEST >)}
(a8) and includes serious sickness by AIDS	{(SEM, <MEDICAL\DISEASE>)}	{(CON, <MEDICAL\DISEASE>)}

TABLE 2. Rules of the first phase for text B in Figure 2

Expressions	*Multi-Attribute rules*	*Output of the first phase*
(b1) "Dr. Sara White"	{(SEM, <OCCUPATION\PROFFESI ONAL>)} {(CAT, NOUN)}	{(CON, <NAME\HUMAN>)}
(b2) November 14th, 2009	{(SEM, <DATE>)}	{(CON, <DATE >)}
(b3) Director of Johnson Hospital	{(SEM, <JOB\TITLE>)} {(SEM, <NAME\HOSPITAL>)}	{(CON, <JOB\TITLE>)}
(b4) Site News	{(SEM, <HOME PAGE>)}	{(CON, < HOME PAGE >)}
(b5) Blood type is Rh-.	{(SEM, <MEDICAL\BLOODTYPE>)} {(SEM, <BLOODTYPE VALUE>)}	{(CON, <MEDICAL\BLOO DTYPE>)}
(b6) Visit to hospital home page	{(SEM, <INTRODUCTION>)} {(SEM, < HOME PAGE>)}	{(CON, < INTRODUCTION>)}
(b7) if you want to see , it becomes over 40 IU in GPT	{(SEM, <REQUEST>)} ---------- {(SEM, <VALUE>)} {(SEM, <MEDICAL\TEST>)}	{(CON, <REQUEST>)} ---------- {(CON, <MEDICAL\TEST >)}
(b8) and includes serious sickness by AIDS	{(SEM, <MEDICAL\DISEASE>)}	{(CON, <MEDICAL\DISEA SE>)}

TABLE 3. Rules of the second phase for text A in Figure 1

Multi-Attribute rules	*Output of the second phase*
{(CON, <NAME\HUMAN>)} {(CON, <DATE\BIRTHDAY>)} {(CON, < MEDICAL>)}	{(FIX, < MEDICAL>)}
{(CON, <NAME\HUMAN>)} {(CON, <DATE\BIRTHDAY>)} {(CON, <MEDICAL>)}	{(FIX, < MEDICAL>)}
{(CON, <NAME\HUMAN>)} {(CON, <EXAMINATION>)} {(CON, <MEDICAL>)}	{(FIX, < MEDICAL>)}

TABLE 4. Rules of the second phase for text B in Figure 2

Multi-Attribute rules	*Output of the second phase*
{(CON, <NAME\HUMAN>)} {(CON, <DATE\BIRTHDAY>)} {(CON, <HOME PAGE>)}	{(NON, < MEDICAL>)}
{(CON, <NAME\HUMAN>)} {(CON, <DATE\BIRTHDAY>)} {(CON, <INTRODUCTION>)}	{(NON, < MEDICAL>)}
{(CON, <NAME\HUMAN>)} {(CON, <DATE\BIRTHDAY>)} {(CON, < REQUEST>)}	{(NON, < MEDICAL>)}

3.2. **Pattern matching machine.** Multi-attribute matching has been used in many earlier papers. Ando [27,28] proposed a set of matching algorithm with its implementation developed in C programming language. Kadoya [15] used this algorithm for E-mail processing, Kiyoi [29] used his/her approach for medical reports and Yoshinari [26] used his/her approach for emotion analysis. This section depends on the formal definition by Kiyoi [29], and detailed constructions are extended in this section.

In this method, MAPM (multi-attribute pattern-matching) machine takes R as the input and produces matching results as the output corresponding to the rules, assuming that R is a sequence of the input structures. The machine MAPM formally consists of a set of states and each state is represented by a number. Although the matching operation of the machine MAPM is similar to the multi-keyword string pattern-matching method of Aho-Corasick [30], it has the following distinctive features:

goto and output functions

Let T be a set of states and let L be a set of the rule structures R, then the behavior of the machine MAPM is defined by the next two functions:

$*$ **goto function** *goto*: $T \times L \rightarrow T \cup \{fail\}$ where the function goto maps a set of consisting of a state and a rule structure into a state or the message *fail*. A transition label of the *goto* function is extended to a set of notation. Therefore, in the machine MAPM, a confirming transition is decided by the inclusion relationship whether the input structure N includes the rule structure R or not.

$*$ **output function** *output:* T is mapped into the meanings of personal and confidential expressions, or the meanings for context analysis. These meanings are defined for rules. The same set of representation is also able to define the input structures to be matched by the matching rule. N is used as the notation for input structures to distinguish them from R. Matching of the rule structure R and the input structure N are decided by the inclusion relationship such that N includes R ($N \supseteq R$), in order to consider the abstraction of the rule structure. According to this processing, the machine MAPM is also called a set of matching machine.

As discussed in [15,25], the machine MAPM becomes non-deterministic if there are two more labels R and W such that the current input structure N includes both R and W of *goto (s, R)* and *goto (s,W)* for the current state s. The ambiguity can be solved by selecting *goto* (s,R) such that the number of elements in the intersection N and R is larger than that of N and W.

Figures 6 and 7 show the *goto* and *output* functions for rules in Tables 1 and 4, respectively. These functions for Tables 2 and 3 are ignored because they are straightforward.

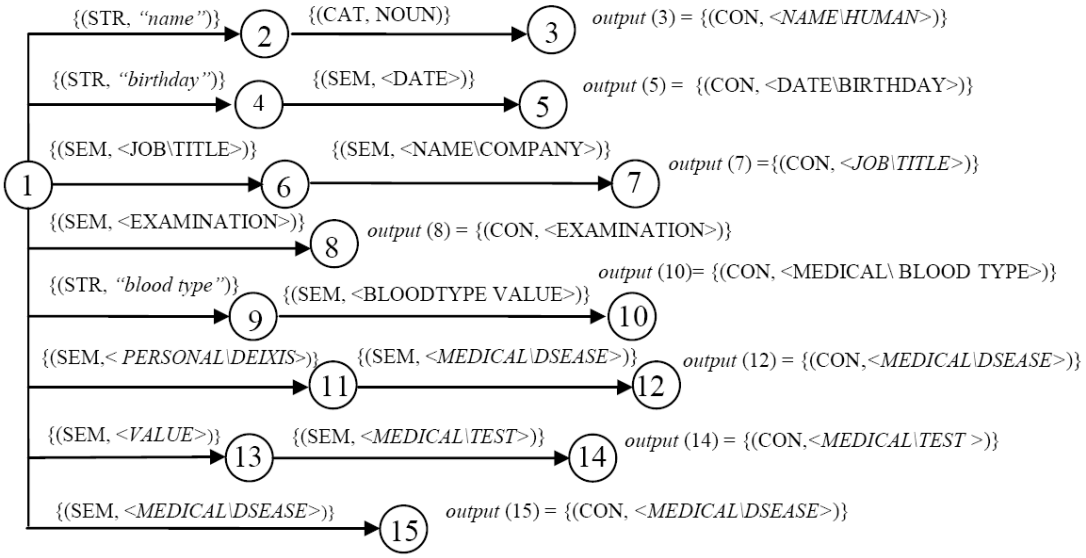


FIGURE 6. *goto* and *output* functions of Table 1

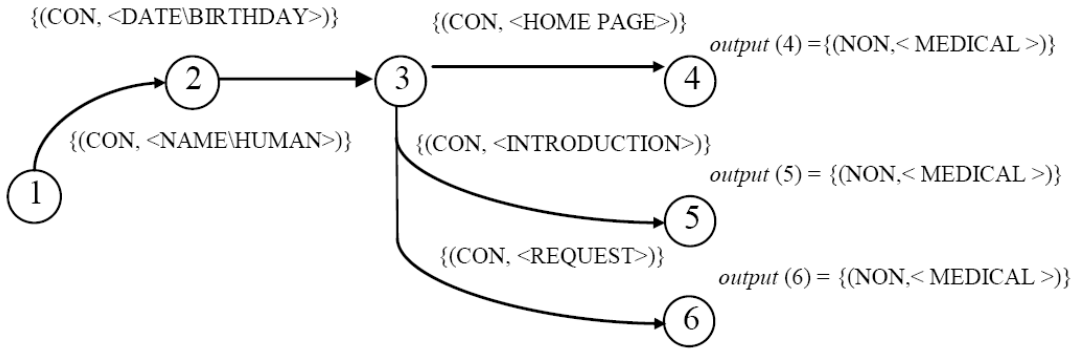


FIGURE 7. *goto* and *output* functions of Table 4

4. **Context Analysis by Multi-attribute Matching.** The context analysis is carried out by a two phase process, where the rules of Tables 1 and 2 are used for the first phase, and those of Tables 3 and 4 are used for the second phase. The following Algorithm 1 is a generalized behavior of the context analysis of the proposed method.

Algorithm 1

Input: A sequence α of input structures is $N_1, N_2, \ldots, N_n$, where each N_i $(0 < i < n+1)$ is an input structure. M is a multi-attribute machine defined by *goto* and *output* functions.

Output:

Method:

(Step 1) Initialization

Initialize STATE *by* 1;
Initialize TEMP *by* {STATE};
/* {} is the empty set */
Initialize LOC *by* 1;

(Step 2) Confirmation of transition

For each STATE in TEMP **do**
  For each R and for NEXT such that *goto* (STATE, R) = NEXT is defined **do**;
  /* Note that undefined *goto* (STATE, R) becomes 0. */
    if INCLUDE (N_{LOC}, R) = true

```
            then append NEXT to NEXTTEMP;
          else if PRIFIX (N_LOC, R) = true
              then append NEXT to NEXTTEMP;
            else if UNKNOWN (N_LOC, R) = true
              then append NEXT to NEXTTEMP;
      /* Note that the total number of elements in NEXTTEMP is restricted
      in the practical system*/
(Step 3) Confirmation of output
For each STATE in NEXTTEMP do
  begin
      if output (STATE) is defined then produce output (STATE);
      if SATE has no transition then remove SATE from NEXTTEMP;
  end;
Increment LOC;
if LOC becomes n+1 then terminate algorithm;
append all elements in NEXTTEMP to TEMP;
Initialize NEXTTEMP by {};
goto (Step 1)
```

(end of Algorithm 1)

```
function INCLUDE (N, R)
/* Transitions by proper inclusion of N and R. */
begin
if N⊇R then return true else return false
end;

function PRIFIX (N, R)
/* Transitions by prefix matching for elements of N and R */
Begin
For element (x, y) in N and for element (x', y') in R,
if y is the prefix of y' then return true else return false
end;

function UNKNOWN (N, R)
/* Transitions by prefix matching for concepts of N and R */
begin
    if N includes (CAT, UNKNOWN) then
        begin
            Set {(CAT, UNKNOWN)} to R;
            return INCLUDE (N, R)
        end;
end;
```

The presented machine M is nondeterministic and TEMP stores the current active states. In Step 1, TEMT has the initial state 1. Step 2 confirms all possible transitions for all sates in TEMP and the next state NEXT is appended to NEXTTEMP. In Step 3, the output function is confirmed and redundant states with no transitions are removed from NEXTTEMP to be merged into TEMP. Although the above redundant states should be checked when NEXT is appended to NEXTTEMP in Step 2 in the time efficiency, Algorithm 1 describes them in Step 3 for readability.

Step 2 performs three types of matching processes. INCLUDE (N, R) is based on $N \supseteq R$ condition, but PRIFIX (N, R) matches the prefix instead of the original element in R and it enables us to extend possible matching. Suppose that R has the original element (SEM,

<NAME\HUMAN>) and N has the prefix element (SEM, <NAME>). Then, INCLUDE (N, R) becomes false, but PRIFIX (N, R) is true. The prefix matching is practical in the case that the concrete value can not be determined in the input. UNKNOWN (N, R) is relation to the robustness for unknown or new expressions. Suppose that N has (CAT, UNKNOWN) when the input has unknown expressions. Then, R is replaced by {(CAT, UNKNOWN)} and INCLUDE (N, R) is invoked. Therefore, transitions are always success. Note that the total number of replacements is restricted in the practical system because of time efficiency.

Consider the basic matching process by INCLUDE (N, R) for *(a6)* and *(a7)* in Figure 1.

Table 5 shows the flow of the first phase matching for *(a6)* and *(a7)*, where TEMP is neglected. For this input, all N_{LOC} can includes R, that INCLUDE (N_{LOC},R) = true, then it is clear that *output* (12) = {(CON, <MEDICAL\DISEASE>)} and *output* (14) = {(CON, <MEDICAL\TEST>)} are obtained from Table 5.

TABLE 5. Examples of the first phase of matching process

STATE	N_{LOC}	*R*	*goto*(STATE,R) *output*(SATAE)
1	N_1 = {(STR, *"she"*), (CAT, NOUN), (SEM, <PERSON\DEIXIS>)}	{(SEM, <PERSON\DEIXIS>)}	4
11	N_2 = {(STR, *"acute liver inflammation"*), (CAT, NOUN), (SEM, <MEDICAL\DISEASE>)}	{(SEM, <MEDICAL\DISEASE>)}	12
12	N_3 = {(STR, *"40IU"*), (CAT, NOUN), (SEM, <VALUE>)}	*goto* function becomes failure	*output*(12) = {(CON, <MEDICAL\DISEASE>)}
1	N_3 = {(STR, *"40IU"*), (CAT, NOUN), (SEM, <VALUE>)}	{(SEM, <VALUE>)}	13
13	N_4 = {(STR, *"GTP"*), (CAT, NOUN), (SEM, <MEDICAL\TEST>)}	{(CAT, NAME\HOSPITAL)}	14 *output*(14) = {(CON,<MEDICAL\TEST>)}
14	LOC = n+1 =5		Terminate algorithm 1

Suppose that N_3 = {(STR, *"40abc"*), (CAT, UNKNOWN)} for unknown expressions "40abc". The function INCLUDE and PREFIX are not success and the invoked function UNKNOWN (N_3, R) is success because R is changed by {(CAT, UNKNOWN)}. Although the function UNKNOWN produces many accessible transitions, it is a very practical scheme with robustness because it is easy to restrict the upper bound of possible transitions to TEMP. Of course, we can say that the less transitions with UNKNOWN become appropriate results.

For the following outputs (inputs of the second phase) by the first phase for Text B of Figure 2, consider the results of the second phase matching by rules of Table 2.

1) N_1 = {(CON, <NAME\HUMAN>)},
   goto (1, {(CON, <NAME\HUMAN>)} = 2, TEMP = {1, 2}
2) N_2 = {(CON, <DATE >)},
   INCLUDE(N_2, R) becomes false, but PRIFIX(N_2, R) = true because <DATE> in N_2 is the prefix of <DATE\BIRTHDAY> in R = {(CON, <DATE\BIRTHDAY>)}.
   NEXT becomes *goto* (2, R) = 3, TEMP = {1, 2, 3}
3) N_3 = {(CON, < TITLE\JOB>)}, *Non transition.*
4) N_4 = {(CON, <HOME PAGE>)},
   goto (3, {(CON, <HOME PAGE>)} = 4, TEMP = {1, 2, 3, 4}
   output (4) = {(NON, <MEDICAL>)}

5) N_5 = {(CON, <MEDICAL\BLOOD TYPE>)}, *Non transition.*
6) N_6 = {(CON, <INTRODUCTION>)},
goto (3, {(CON, <INTRODUCTION>)} = 5, TEMP = $\{1, 2, 3, 4, 5\}$
output (5) = {(NON, <MEDICAL>)}
7) N_7 = {(CON, <REQUEST>)},
goto (3, {(CON, <REQUEST>)}) = 6, TEMP = $\{1, 2, 3, 4, 5, 6\}$
output (6) = {(NON, <MEDICAL>)}
8) N_8 = {(CON, <MEDICAL\TEST>)}, *Non transition.*
9) N_9 = {(CON, <MEDICAL\DISEASE>)}, *Non transition.*

Although the multi-attribute rules of Table 3 determine (FIX, <MEDICAL>), the above output specifies (NON, <MEDICAL>). Therefore, the final results of Text B become non-PC by the NEG concept (NON, <MEDICAL>).

5. **Experimental Evaluations.** In this study, PM is a new detection scheme based on context analysis to improve the accuracy rate by reducing the extra-detections for non-PC texts. In order to estimate the efficiency of the presented method, TM is introduced as a scheme based on sentence analysis or non-context analysis.

This section evaluates the presented extraction method for personal and confidential information. Two criterions called precision and recall are used to evaluate the extraction accuracy for PC information in the presented and traditional methods as follows:

Let TNE be the total number of extracted PC information, let TNCE be the total number of correct extracted PC information, and let NC be the number of correct PC information. Then, Precision (%) is defined by (NC/TNE)*100 and Recall (%) is (NC/TNCE)*100.

5.1. **Experimental data and results.** It is difficult to prepare texts with actual PC expressions, so the following experimental data have been collected:

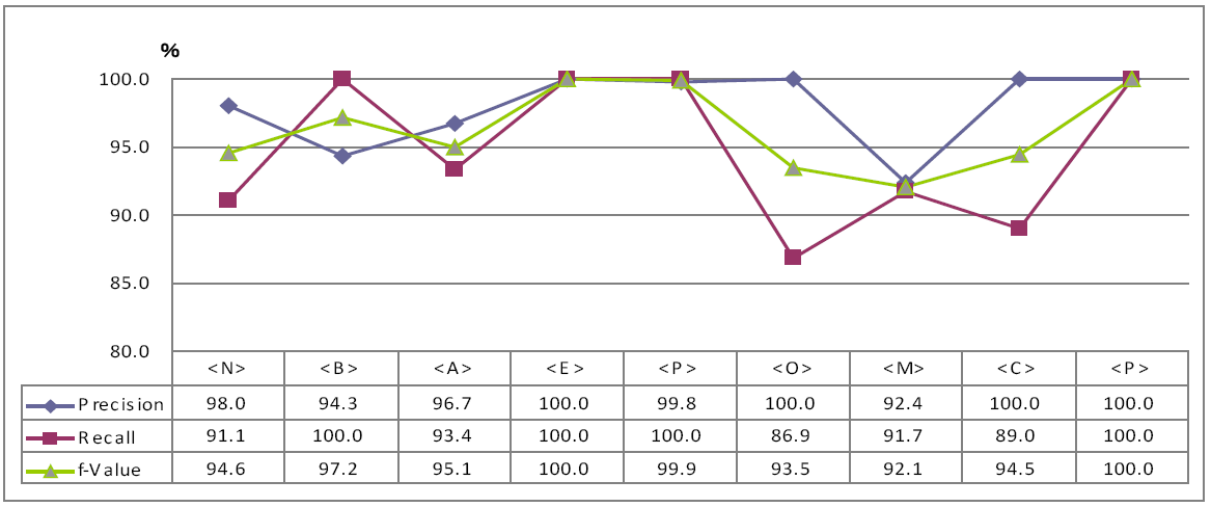
1) Electronic medical recoding (MR) texts
MR includes <MEDICAL\BODY CONDITION> and <MEDICAL\DISRASE> [29], but personal information was deleted.
2) Organization (ORG) texts
ORG includes <NAME>, <ADDRESS>, <E-MAIL>, <PHONE NUMBER> and <FINACIAL\STOCK>) [31].
3) Personal (PER) texts
PER includes <NAME>, <BIRTHDAY>, <E-MAIL> and <PHONE NUMBER>, <ORGANIZATION>, <NAME\LOGIN>, <PASSWARD> [32].
4) Non-PC (NPC) texts
NPC texts include NEG concepts <NAME>, <INTRODUCTION>, <REQUEST>, <QUESTION>, <EMOTION>, <REPLY>, <INVITATION>, <ENCOURAG EMENT>, <ASSUMPTION> [23].

Table 6 represents the total number of selected experimental data, where <MEDICAL\ BODY CONDITION> and <MEDICAL\DISRASE> are merged into <MEDICAL>.

For the experimental data shown in Table 6, Figure 8 shows precision and recall of detecting each concept from PC texts, where <N> is <NAME>, <B> is <BIRTHDAY>, <A> is <ADDRESS>, <E> is <E-MAIL>, <P> is <Phone NUMBER>, <O> is <ORGANIZATION>, <M> is <MEDICAL>, <E> is <E-MAIL> and <P> is <PATE NT>. Figure 9 shows precision and recall of detecting each concept from NON-PC expressions, where <IN> is <INTRODUCTION>, <RQ> is <REQUEST>, <Q> is <QUESTION>, <EM> is <EMOTION>, <RP> is <REPLY>, <IV> is <INVITATIO

TABLE 6. Information of experimental data

Concepts	MR	ORG	PER	NPC
<NAME>	0	6,000	1,000,000	4,316
<BIRTHDAY>	0	0	1,000,000	0
<ADDRESS>	0	6,000	1,000,000	0
<E-MAIL>	0	6,000	1,000,000	0
<PHONE NUMBER>	0	6,000	1,000,000	0
<ORGANIZATION>	0	6,000	0	0
<MEDICAL\DISEASE>	2,000,000	0	0	0
<INTRODUCTION>	0	0	0	1,460
<REQUEST>	0	0	0	738
<QUESTION>	0	0	0	108
<EMOTION>	0	0	0	259
<REPLY>	0	0	0	42
<INVITATION>	0	0	0	44
<ENCOURAGEMENT>	0	0	0	82
<ASSUMPTION>	0	0	0	331



	<N>	<B>	<A>	<E>	<P>	<O>	<M>	<C>	<P>
Precision	98.0	94.3	96.7	100.0	99.8	100.0	92.4	100.0	100.0
Recall	91.1	100.0	93.4	100.0	100.0	86.9	91.7	89.0	100.0
f-Value	94.6	97.2	95.1	100.0	99.9	93.5	92.1	94.5	100.0

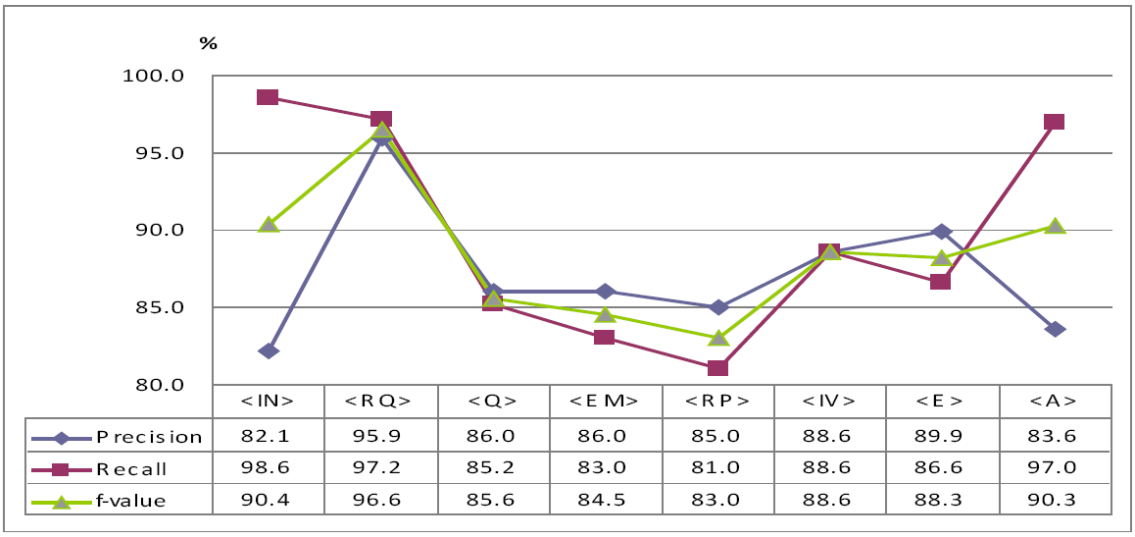
FIGURE 8. Results of precision and recall for each concept for PC expressions

N>, <E> is <ENCOURAGE> and <A> is <ASSUMPTION>. Figure 8 depends on the results in [25].

From results of Figures 8 and 9, it is clear that the accuracy for each concept is very high.

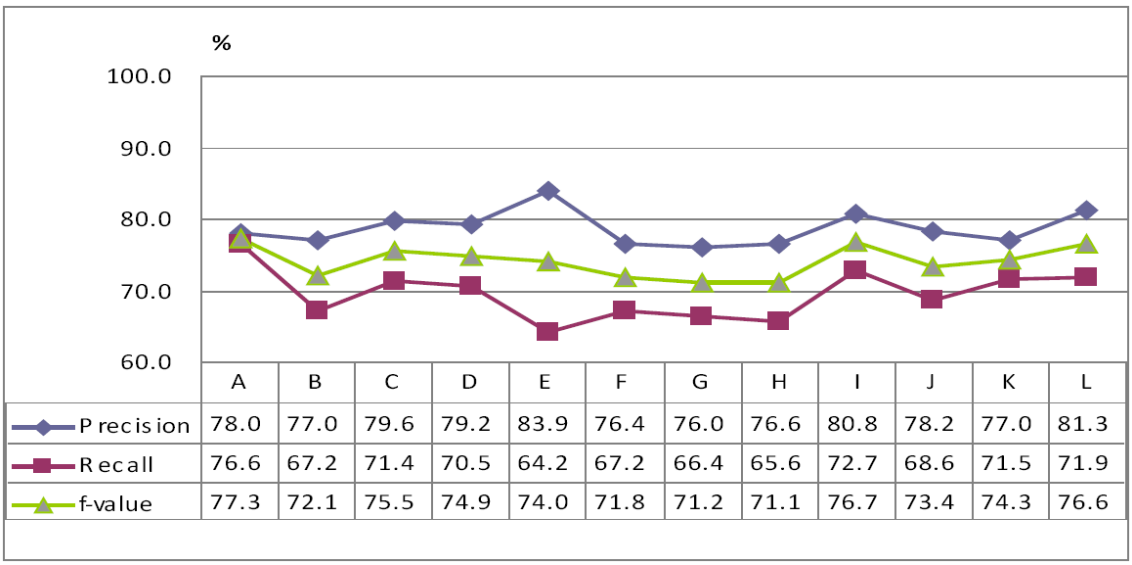
From Figures 8 and 9, the accuracy of neglecting non-PC texts including both PC expressions and NEG expressions is estimated which confirmed that the accuracy of the presented method for PC information is practical. Therefore, the evaluation has been confirmed by combining the results of the above PC expressions (<NAME(N)>, <ADDRESS(A)>, <MEDICAL(M)>) in Figure 8 with the following text groups including two types of NEG concepts in Figure 9.

Group A: <IN> and <RQ> Group B: <IN> and <Q>
Group C: <RQ> and <Q> Group D: <RQ> and <EM>
Group E: <EM> and <RP> Group F: <EM> and <IV>
Group G: <PR> and <E> Group H: <PR> and <A>



	<IN>	<RQ>	<Q>	<EM>	<RP>	<IV>	<E>	<A>
Precision	82.1	95.9	86.0	86.0	85.0	88.6	89.9	83.6
Recall	98.6	97.2	85.2	83.0	81.0	88.6	86.6	97.0
f-value	90.4	96.6	85.6	84.5	83.0	88.6	88.3	90.3

FIGURE 9. Experimental results for NEG concepts for non-PC expressions



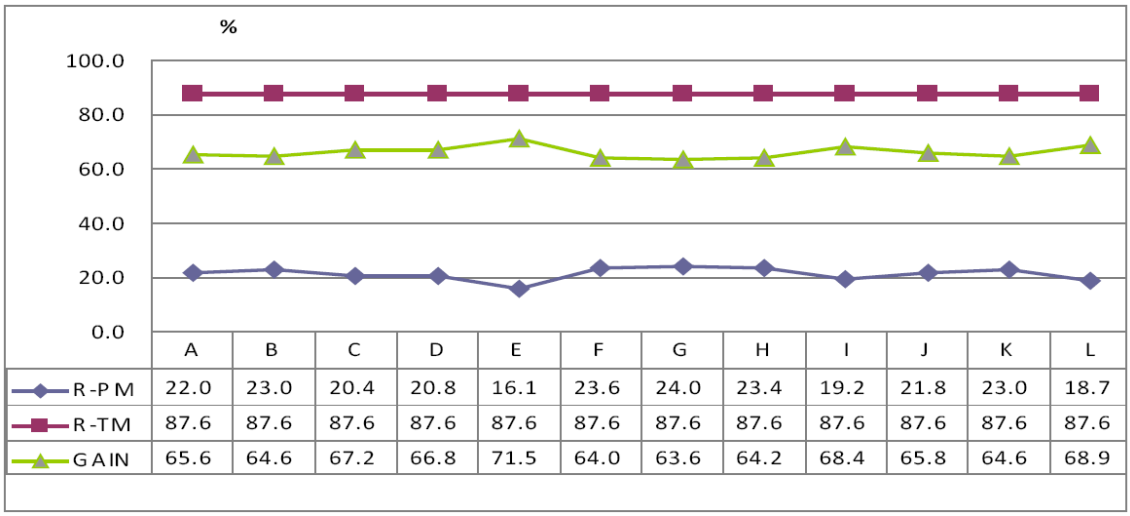
	A	B	C	D	E	F	G	H	I	J	K	L
Precision	78.0	77.0	79.6	79.2	83.9	76.4	76.0	76.6	80.8	78.2	77.0	81.3
Recall	76.6	67.2	71.4	70.5	64.2	67.2	66.4	65.6	72.7	68.6	71.5	71.9
f-value	77.3	72.1	75.5	74.9	74.0	71.8	71.2	71.1	76.7	73.4	74.3	76.6

FIGURE 10. Experimental results for non-PC texts

Group I: <IV> and <E> Group J: <IV> and <IN>
Group K: <E> and <IN> Group L: <E> and <RQ>

Figure 10 shows experimental results from the presented context analysis for non-PC texts. Consider group A (<IN> and <RQ>) to explain results in Figure 10. First of all, the basic rate 87.6% to detect PC expressions is obtained by multiplying 98.0% of <N>, 96.7% of <A> and 92.4% of <M> in Figure 8. Second, the rates 71.9% of <IN> and 84.0% of <RQ> are calculated by multiplying 87.6 for 82.1% of <IN> and 95.9% of <RQ> as in Figure 9, respectively. Then, precision value 78.0% is obtained by the averages 71.9% of <IN> and 84.0% of <RQ>.

Let R-PM and R-TM be the rate of extra-detection for the presented method and the traditional method, respectively. Let GAIN be the improvement rate calculated by subtracting R-PM from R-TM. Figure 11 shows a comparison of the PM with TM for non-PC expressions. The result 22.0% of R-PM for group A in Figure 11 is obtained by subtracting the precision 78.0% in Figure 10 from 100%. While the result 87.6% of R-TM



	A	B	C	D	E	F	G	H	I	J	K	L
R-PM	22.0	23.0	20.4	20.8	16.1	23.6	24.0	23.4	19.2	21.8	23.0	18.7
R-TM	87.6	87.6	87.6	87.6	87.6	87.6	87.6	87.6	87.6	87.6	87.6	87.6
GAIN	65.6	64.6	67.2	66.8	71.5	64.0	63.6	64.2	68.4	65.8	64.6	68.9

FIGURE 11. The comparison between the presented context analysis and traditional methods for non-PC texts

is obtained by multiplying 98.0% of <N>, 96.7% of <A> and 92.4% of <M> in Figure 8.

From results of Figure 11, it turns out that GAIN of the presented method can achieve average 66.5%. Suppose that the number of target texts is 100,000. Then, although there are 87,600 miss-detection texts in the traditional method, the number can be reduced to 21,300 through the presented method. That is to say, the number of persons checking these texts can be reduced from 37 days to 9 days if one person confirms 2,367 texts per day, this mean the human effort is reducing by about 80% than by using the traditional method.

The detected results can also use SVM schemes as the learning futures. However, the extending detecting knowledge of SVM needs to prepare a lot of correct training data. That is to say, the accuracy of detecting is relating to the accuracy of features of SVM and the presented method can co-operate SVM.

Average analysis time is very practical because it is about 33 ms for 6 KB plain text and about 83 ms for 100 KB HTML texts on Intel CPU E5440 (2.85 Hz).

Introduction table of famous person and economy news with company reports have many PC expressions, but they are non-PC texts. The presented method is weak for these texts, so the remaining problem is to study solutions by filed recognition techniques [33-36].

6. **Conclusions.** This paper has presented a method of extracting PC information by using context analysis. In the presented scheme, rule based knowledge has been introduced and two-phase process with context processing has been proposed. The traditional methods depend on words or a sequence of words without considering the context, and many irrelevant candidates of possible PC information are extracted. The presented method can solve these problems by introducing NEG concepts, context analysis of a two-phase process using multi-attribute matching. From experimental results for a large amount of text data, it has been verified that the presented method could improve the low rate of precision for non-PC texts. Therefore, we can say that the presented method is a very useful approach for personal and confidential information filtering services expressions.

Future work could focus on using more categories in personal and confidential information expressions to improve the presented method.

Acknowledgments. This research has been partially supported by the Ministry of Education, Culture, Sports, Science and Technology in Japan, Grants-in-Aid for Scientific Research (A), 22240020.

REFERENCES

[1] *MCI*, http://news.cnet.com/Credit-card-breach-exposes-40-million-accounts/2100-1029_3-5751886.html.
[2] *ATMs*, http://news.cnet.com/8301-10784_3-9982500-7.html.
[3] *HACK*, http://news.cnet.com/8301-1009_3-10372523-83.html.
[4] *DPD*, http://en.wikipedia.org/wiki/Data_Protection_Directive.
[5] *PIPL*, http://www.ibls.com/internet_law_news_portal_view.aspx?s=latestnews&id=2242.
[6] *HIPAA*, http://en.wikipedia.org/wiki/Information_privacy.
[7] M.-C. Landau, F. Sillion and F. Vichot, Exoseme: A thematic document filtering system, *Intelligence Artificial Journal*, 1993.
[8] M. Larry and Y. Malik, One-class SVMs for document classification, *Journal of Machine Learning Research*, pp.139-154, 2001.
[9] K. Yoohwan, C. Wing, C. Mooi and H. Jonathan, PacketScore: A statistics-based packet filtering scheme against distributed denial-of-service attacks, *IEEE Trans. on Dependable and Secure Computing*, vol.3, no.2, pp.141-155, 2006.
[10] W. Lee, S. Lee, S. Chung and D. An, Harmful contents classification using the harmful word filtering and SVM, *Proc. of the 7th International Conference on Computational Science, Part III*, Beijing, China, pp.18-25, 2007.
[11] G. Bruno, B. Stephan and L. Philippe, Combining classifiers for harmful document filtering, *Proc. of RIAO International Conference*, Avignon, France, pp.173-185, 2004.
[12] W. Francis, V. Frantz and D. Bruno, Automatic processing of proper names in texts, *Proc. of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, University College Dublin, Belfield, Dublin, Ireland, pp.23-30, 1995.
[13] E. Richard, A framework for named entity recognition in the open domain, *Proc. of the Recent Advances in Natural Language Processing*, Borovets, Bulgaria, pp.137-144, 2003.
[14] L. M. Manevitz and M. Yousef, One-class SVMs for document classification, *Journal of Machine Learning Research*, vol.2, pp.139-154, 2001.
[15] K. Kadoya, M. Fuketa, E.-S. Atlam and J. Aoe, An efficient e-mail filtering using time priority measurement, *International Journal of Information Science*, vol.166, pp.213-229, 2004.
[16] *SR*, http://en.wikipedia.org/wiki/Street_or_road_name.
[17] *LCC*, http://en.wikipedia.org/wiki/List_of_country_calling_codes.
[18] *UCSD*, http://www.bennetyee.org/ucsd-pages/area.html.
[19] *BT*, http://en.wikipedia.org/wiki/Blood_type.
[20] *ID*, http://en.wikipedia.org/wiki/Infectious_disease.
[21] *GE*, http://en.wikipedia.org/wiki/Grade_(education).
[22] *CSC*, http://www.exportbureau.com/check_swift_code.html.
[23] *IBAN*, http://en.wikipedia.org/wiki/International_Bank_Account_Number.
[24] *TS*, http://en.wikipedia.org/wiki/Ticker_symbol.
[25] Y. Kadoya, K. Morita, M. Fuketa, O. Masaki, E.-S. Atlam, S. Toru and J. Aoe, A sentence classification technique by using intention association expressions, *Computer Mathematics*, vol.82, no.7, pp.777-792, 2005.
[26] T. Yoshinari, E.-S. Atlam, K. Morita, K. Kiyoi and J. Aoe, Automatic acquisition for sensibility knowledge using co-occurrence relation, *International Journal of Computer Applications in Technology*, vol.33, no.2/3, pp.218-225, 2008.
[27] K. Ando, T. Kinoshita, M. Shishibori and J. Aoe, An improvement of the Aho-Corasick machine, *International Journal of Information Sciences*, vol.111, pp.139-151, 1998.
[28] K. Ando, S. Mizobuchi, M. Shishibori and J. Aoe, Efficient multi-attribute pattern matching, *International Journal of Computer Mathematics*, vol.66, no.1-2, pp.21-38, 1998.
[29] K. Kiyoi, E.-S. Atlam, M. Fuketa, Y. Tomoko and J. Aoe, A method for extracting knowledge from medical texts including numerical representation, *International Journal of Computer Applications in Technology*, vol.33, no.2/3, pp.226-236, 2008.
[30] A. V. Aho and M. J. Corasick, Efficient string matching: An aid to bibliographic search, *Communications of the ACM*, vol.18, no.6, pp.333-340, 1975.

[31] *ShikiHou 29 (Company Information)*, Toyokeizai, 2007.
[32] *Pseudo Personal Data*, People to People Communications, 2007.
[33] E.-S. Atlam, M. Fuketa, K. Morita and J. Aoe, Documents similarity measurement using field association terms, *International Journal of Information Processing and Management*, vol.39, no.6, pp.809-824, 2003.
[34] E.-S. Atlam, G. Elmarhomy, M. Fuketa, K. Morita and J. Aoe, Automatic building of new field association word candidates using search engine, *Information Processing & Management Journal*, vol.42, no.4, pp.951-962, 2006.
[35] M. Fuketa, S. Lee, T. Tsuji, M. Okada and J. Aoe, A document classification method by using field association words, *International Journal of Information Sciences*, vol.126, no.1, pp.57-70, 2000.
[36] M. Fuketa, K. Kiyoi, E.-S. Atlam, K. Tsutomu, K. Morita, K. Shinkaku and J. Aoe, A method of extracting and evaluating good and bad reputations for NL expressions, *Information Technology & Decision Making*, vol.4, no.2, pp.77-196, 2005.