

A CONTEXT ANALYSIS SCHEME OF DETECTING PERSONAL AND CONFIDENTIAL INFORMATION

TOSHIHIRO SATOMI¹, ATLAM EL-SAYED^{1,2}, KAZUHIRO MORITA¹
MASAO FUKETA¹ AND JUN-ICHI AOE¹

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

²Department 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: